



Durham E-Theses

Views from the students' desks: How students experience and comprehend demand and difficulty in GCSE mathematics examination questions

FOWLER, ANDREW, THOMAS

How to cite:

FOWLER, ANDREW, THOMAS (2023) *Views from the students' desks: How students experience and comprehend demand and difficulty in GCSE mathematics examination questions*, Durham theses, Durham University. Available at Durham E-Theses Online: <http://etheses.dur.ac.uk/15238/>

Use policy

The full-text may be used and/or reproduced, and given to third parties in any format or medium, without prior permission or charge, for personal research or study, educational, or not-for-profit purposes provided that:

- a full bibliographic reference is made to the original source
- a [link](#) is made to the metadata record in Durham E-Theses
- the full-text is not changed in any way

The full-text must not be sold in any format or medium without the formal permission of the copyright holders.

Please consult the [full Durham E-Theses policy](#) for further details.

Views from the students' desks:

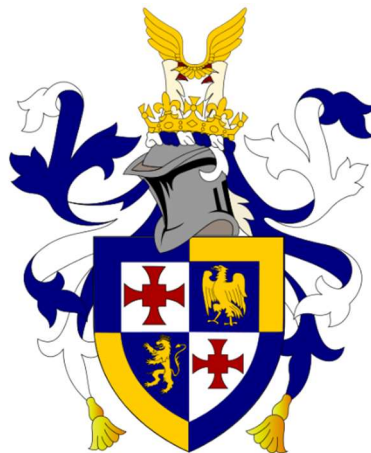
How students experience and comprehend demand and difficulty in GCSE mathematics examination questions

Andrew Thomas Fowler

Thesis presented for the degree of

Doctor of Philosophy

at the School of Education, Durham University



St John's College

Durham University

United Kingdom

2023

Abstract

This study investigates students' understanding and experience of factors that create "demand" and "difficulty" in GCSE mathematics examination questions.

Around 600,000 students in England take GCSE examinations each year. The results of these high-stakes assessments affect students' future prospects, as well as schools' status and recruitment. The performance of the examination system is, therefore, highly significant, but the effective working of its assessment methods lacks systematic academic scrutiny.

Examiners manipulate factors relating to the "demand" (i.e. cognitive load) of examination questions, and these questions are experienced at different levels of "difficulty" by students. If the link between "demand" and "difficulty" cannot accurately be predicted by examiners, this poses a threat to the validity of inferences made from examination results. Existent research into demand and difficulty in examination questions has predominantly focused on the work of examiners: students' voices have not been heard.

In this study, questionnaires and focus group interviews gathered the views of 224 secondary school GCSE mathematics students in 5 comprehensive schools in North East England. Reflexive thematic analysis (Braun and Clarke, 2022) was used to investigate students' responses and their inferences in relation to learning theories including cognitive load theory and taxonomies of learning.

The voices of students in this study reveal insights into question demands and difficulty. Students discuss recall, reasoning and application of knowledge. Many students associate question length with difficulty. Question clarity inspires student confidence. Context in a question introduces an unreliable element, motivating some and confusing others.

Students' insights have implications for examiners, teachers, and students. The thesis concludes that there are compelling reasons, in terms of teaching and learning, improving question design, validity, and public confidence in the examination system, to listen to students' views.

Key terms: assessment; public examinations; validity; student voice; learning theories; motivation; reflexive thematic analysis.

(298 words)

This page has been left intentionally blank

Table of Contents

Abstract.....	1
Table of Contents.....	3
List of Figures.....	7
List of Tables.....	8
Declaration.....	9
Statement of Copyright.....	9
Acknowledgements.....	11
Dedication.....	13
Foreword.....	15
Preface.....	17
Journey to a foreign land.....	21
Tone of voice.....	25
Chapter 1: Introduction – The idea of a journey.....	27
1.1 Context: the GCSE examination system in England, Wales, and Northern Ireland.....	29
1.2 Socio-political aspects of the current high-stakes assessment system.....	29
1.3 How examination questions should work in a simplified system.....	32
1.4 Rationale for this study.....	40
1.5 Researcher positioning.....	42
1.6 Research aim and question.....	46
1.7 Impact of the COVID-19 pandemic on this study.....	47
1.8 Structure of Thesis.....	48
Chapter 2: The concepts of “Demand” and “Difficulty” – Learning the language.....	51
2.1 Demand in examination board literature.....	53
2.2 Difficulty in examination board literature.....	58
2.3 Demand and Difficulty: Conclusion.....	71
Chapter 3: Literature Review – Mapping the known world.....	73
3.1 Working memory and cognitive load theory, and their relevance to public examinations.....	76
3.1.1 Working memory.....	76
3.1.2 Cognitive load theory.....	79
3.1.3 Optimising cognitive load, and Vygotsky’s Zone of Proximal Development.....	92
3.1.4 Critical engagement with Cognitive Load Theory.....	96
3.2 Taxonomies of learning, student responses and question demands.....	101
3.2.1 Bloom’s Taxonomy.....	101

3.2.2 Marzano and Kendall: A New Taxonomy	108
3.2.3 Noel Burch Competence Model	112
3.2.4 CRAS Scales of Demands	114
3.3 Validity considerations underpinning high-stakes assessments	119
3.3.1 Validity in relation to GCSE examinations	123
3.4 Student voice	128
3.4.1 Recent secondary school student voice studies – an evaluation	142
3.4.2 Student Voice in Higher Education and in Primary Schools	147
3.5 Literature Review Conclusion	151
Chapter 4: Methods – Planning my immersive journey	155
4.1 Philosophical location of the current study	156
4.2 Ensuring validity	160
4.3 Ethical considerations	166
4.4 Questionnaire design	167
4.5 Study samples	173
4.6 Focus Groups	178
4.7 Data analysis techniques: descriptive and inferential statistics	181
4.8 Data analysis techniques: reflexive thematic analysis	183
4.9 Summary of the philosophical approach and methods in this study	195
Chapter 5: Pilot Study – Embarking on my first proper trip	197
5.1 Data analysis: descriptive statistics	197
5.2 Data analysis: inferential statistics	203
5.3 Discussion of pilot study survey questions	205
Question 1	208
Question 2	213
Question 3	215
Question 4	219
Question 5	220
Question 6	223
5.4 Reflexive thematic analysis	226
5.4.1 “Looks Easy” theme	228
5.4.2 “Recognising Steps” theme	234
5.5 Evaluation of Pilot Study – informing the Main Study	240
5.6 Conclusion of the Pilot Study	242
Chapter 6: Main Study – Exploring the new world in more detail	244
6.1 Data analysis: descriptive and inferential statistics	246
6.1.1 Sample size and gender distribution	246

6.1.2 Estimates of difficulty	247
6.1.3 Students' marks for GCSE Mathematics questions.....	248
6.1.4 Correlations.....	253
6.2 Reflexive thematic analysis	254
6.3 "Steps and Methods" theme	256
6.3.1 Attention	259
6.3.2 Structure	262
6.4 "Wording and Clarity" Theme.....	272
6.4.1 Interpretation: words vs numbers	273
6.4.2 Contexts	278
6.5 "Memory, Practice and Familiarity" theme	282
6.5.1 Memory.....	283
6.5.2 Practice and familiarity	286
6.6 "Motivation" theme.....	290
6.7 Conclusion of the Main Study	294
Chapter 7: Discussion – Telling the adventurous story	295
7.1 The 'Big Q' approach – qualitative methods within a qualitative framework.....	295
7.2 Revisiting validity	297
7.3 Students' experiences of fairness in high-stakes assessments.....	298
7.4 Recapping the conceptual model relating demand and difficulty to cognitive load.....	300
7.5 Understanding demand and difficulty	300
7.6 Hearing and understanding students' voices.....	304
7.7 How students experience and comprehend demand and difficulty in examination questions.....	305
7.8 Understanding demand	309
7.9 Unexpected demand and difficulty.....	322
7.9.1 Communication failure, distracting information and context	322
7.9.2 Length of question and the amount of reading required	326
7.9.3 Examination stress	330
7.10 Self-efficacy, student motivation, and fear.....	332
7.11 Comparison of examiners' and students' views	336
Chapter 8: Conclusion – Reflections on the journey	340
8.1 A better understanding of what is meant by "demand" and "difficulty"	341
8.2 Rebalancing the relationship between examiners' and students' perspectives	343
8.3 Hearing students' voices and investigating their understanding of demand and difficulty in examination questions.....	350
8.4 Recommendations, applications, and contributions to enhancing academic knowledge	355

8.5 Limitations of this study	362
8.6 Avenues of future research	364
8.7 Concluding remarks	364
Postscript	367
References	369
Appendices	393
Appendix A: Pilot Study – Information Sheet, Consent Form and Questionnaire	393
Appendix B: Main Study – Information Sheet and Consent Form	405
Appendix C: Main Study – Mathematics GCSE Questions.....	409
Appendix D: Main Study – Questionnaire	413
Appendix E: Focus Group 1 Transcript	415
Appendix F: Focus Group 2 Transcript	427

List of Figures

Figure 1 - A Simplified Examination System	34
Figure 2 - Idealised Relationship between Theory and Assessment Design.....	53
Figure 3 - Relative Difficulty of Questions in GCSE Mathematics.....	63
Figure 4 - How and where the Different Theoretical Frameworks and Conceptual Models operate in the Present Study	75
Figure 5 - Diagram of Human Memory	77
Figure 6 - Anatomy of Cognitive Load.....	85
Figure 7 - Abstract Mathematics Problem.....	87
Figure 8 - Mathematics Problem with a Structured Method	88
Figure 9 - Mathematics Problem, with Step-by-step Approach	89
Figure 10 - Relationship between Intrinsic Load and a Learner's Expertise.....	91
Figure 11 - Vygotsky's Zone of Proximal Development	93
Figure 12 - Bloom's Taxonomy: 1956 Original and Anderson et al.'s 2001 Revision	103
Figure 13 - Lemov: 'Bloom's Delivery Service'	104
Figure 14 - Bloom's Taxonomy (Revised): The Knowledge Dimension.....	105
Figure 15 - Marzano and Kendall: Model of Behaviour	110
Figure 16 - Marzano and Kendall: Levels within the Cognitive System	111
Figure 17 - Noel Burch Competence Model	112
Figure 18 - Three-fold Classification of Student Voice	134
Figure 19 - Relationship between Ontology, Epistemology and Methods in this Study	156
Figure 20 - Representation of Grades in Mathematics GCSE Foundation and Higher Tiers	169
Figure 21 - Data Collection Flowchart for Pilot Study Groups.....	176
Figure 22 - Example of Main Study Student Questionnaire, with Highlights	190
Figure 23 - Extract from Coding Sheet for "Memory, Practice, Familiarity"	192
Figure 24 - Example of Visual Arrangement of Student Comments from Focus Groups	193
Figure 25 - Mapping Developing Themes	194
Figure 26 - Pilot Study: Proportion of Students gaining Full or Zero Marks.....	201
Figure 27 - Line Graph Showing Distribution of Estimates between Groups P and Q.....	204
Figure 28 - Bubble Chart Showing Frequency of Estimates and Marks for Question 4.....	219
Figure 29 - Bubble Chart Showing Frequency of Estimates and Marks for Question 6.....	224
Figure 30 - "Looks Easy" Theme	228
Figure 31 - "Looks Easy" Theme: Judgement and Thinking Process.....	232
Figure 32 - "Recognising Steps" Theme.....	235
Figure 33 - Mapping Themes from the Main Study.....	255
Figure 34 - "Steps and Methods" Theme.....	256
Figure 35 - Ratio Method from Student R34M.....	265
Figure 36 - Tree Diagram from Student R51M	266
Figure 37 - Table Method from Student R36M	267
Figure 38 - Ratio Method from Student R31M.....	269
Figure 39 - "Wording and Clarity" Theme	272
Figure 40 - Conceptual Model of Cognitive Load.....	300
Figure 41 - Conceptual Model, with Sources of Demand and Difficulty.....	308
Figure 42 - Diagram of Human Memory.....	314
Figure 43 - Relationship between Examiners, Researchers and Students	346

List of Tables

Table 1 - CRAS Scales of Demands	115
Table 2 - Steps to Ensure Validity in Qualitative Research in the Current Study.....	162
Table 3 - Validation Principles for Qualitative Study, Related to the present Study	165
Table 4 - Location of GCSE Mathematics Questions	168
Table 5 - Response Rates for Preliminary and Main Studies.....	171
Table 6 - Pilot Study: Timeline of Recruitment	174
Table 7 - Data Collection for Main Study	177
Table 8 - Samples and Gender Distribution	178
Table 9 - Key Differences between Qualitative and Quantitative Paradigms, and their Application to this Study	186
Table 10 - Thematic Analysis in the present Study	188
Table 11 - Pilot Study: Students' Estimates of Difficulty	198
Table 12 - Pilot Study: Students' Marks for GCSE Mathematics Questions.....	200
Table 13 - Pilot Study: Correlations between Students' Difficulty Estimates and Marks	202
Table 14 - Pilot Study: Distribution of Estimates of Difficulty between Groups P and Q	203
Table 15 - CRAS Scales of Demands for Pilot Study GCSE Mathematics Questions.....	207
Table 16 - Evaluation of Pilot Study and Implications for Main Study.....	240
Table 17 - Main Study, Students' Estimates of Difficulty	247
Table 18 - Main Study, Students' Marks	248
Table 19 - Most and Least Difficult Questions	249
Table 20 - Main Study Questionnaire: Reasons Given for Least/Most Difficult Question.....	250
Table 21 - CRAS Scales of Demands for Survey Questions 2 and 7	252
Table 22 - Correlations between Students' Marks and their Estimates of Difficulty.....	253
Table 23 - Group R: Proportion of Response Structures for Question 3.....	270

Declaration

I declare that this dissertation, which I submit for the degree of Doctor of Philosophy, is my own work and has not previously been submitted for a degree at this or any other university.

Statement of Copyright

The copyright of this dissertation rests with the author. No quotation from it should be published with the author's prior written consent, and any information derived from it should be acknowledged.

This page has been left intentionally blank

Acknowledgements

In the Preface, I discuss my motivation for this study. In this section, I want to acknowledge those who provided me with the means and the opportunity to accomplish it.

First, I want to acknowledge the support I have received from Durham University's School of Education in terms of financial support and supervision. This study was fully funded by a research studentship from the School of Education, without which it would have been impossible for me to proceed. My course of study has not been straightforward, however, and it has involved the unusually large number of eight supervisors. I was inspired and motivated by Rob Coe and the late Per Kind, my original supervisors, who exemplified the best traditions of scholarship in the insights, rigour, and encouragement that they showed me. Following the untimely death of Per Kind, and the departure from Durham University of Rob Coe, my supervision passed briefly through the hands of two other members of staff at the School of Education before being taken over by Dimitra Kokotsaki and the late Christine Merrell. I benefited considerably from the experience, imagination, and humanity that Christine brought to her supervision. Following Christine's death, Linda Wang and Xiaofei Qi stepped in. I am grateful to Dimitra for encouraging me, pointing me to the work of Braun and Clarke, and helping me to keep going, and to Linda Wang for her mathematical suggestions.

I am grateful to the governing bodies of the schools I have served as Headteacher and Principal, Dane Court Grammar School and Lord Lawson of Beamish Academy, and particularly to Chris Smith and Guy Currey, Chairs of Governors at Lord Lawson, for supporting this project and allowing me short periods of study leave to write up during the Autumn Term of 2021 and in July 2023; I would like to thank my deputy, Joe Dicocco, for acting as Principal in my absence. I have also enjoyed the support of those headteachers and heads of department in secondary schools in Gateshead who gave permission for data to be collected for the pilot study. This study has focused on the views and experience of students, and I am immensely grateful to the hundreds of students from many different secondary schools who, in differing ways, took part in this study and gave their views, insights, and opinions freely. I hope I have represented their views fully and fairly.

Anthony Parton provided an invaluable perspective in reading an early draft of this study.

Finally, I owe so much to the love, encouragement, and support of my wife, Linda Burton, who believed that this project was possible when I doubted it, and who, through our changes of jobs and house moves, and across nine years, has kept me sufficiently motivated and focused to complete this study. Linda's belief in the importance of listening to the voices of students has driven her own life and career. I hope to have reflected her belief and captured some of the values of her outlook in the study that I now present.

This page has been left intentionally blank

Dedication

This work is dedicated, with love, to my wife, Linda Burton, and with gratitude and respect to the memory of my father, David Michael Fowler.

This page has been left intentionally blank

Foreword

‘When anything important is to be done in the monastery, the abbot shall call the whole community together and himself explain what the business is. And, after hearing the advice of the brothers, let him ponder it and follow what he judges to be the wiser course.

The reason why we say that all should be called to council is this: it is often to the youngest that the Lord reveals new and better solutions.’

Benedict of Nursia (516 AD). The Rule of St Benedict, Chapter 3.

This page has been left intentionally blank

Preface

“I met a traveller from an antique land¹” (Percy Bysshe Shelley – *Ozymandias*, 1818).

Standing outside a school sports hall on a summer’s day as 180 sixteen-year-old students emerged, blinking and stretching, into the bright sunlight, I asked them, as their headteacher, the customary questions – ‘how did it go?’ ‘how did you get on?’ Some were just glad that it (a GCSE mathematics exam) was over; others conducted their own informal post-mortems.

Question 4, in particular, seemed to be the subject of debate: some had ‘found it all right – quite straightforward;’ ‘I knew what to do;’ whereas others had found it much more difficult: ‘I didn’t get it;’ ‘I couldn’t see what they wanted.’

Reflecting upon these different experiences, I wondered: how could one question have been experienced so differently by all these students? It was possible that some students had revised more thoroughly, or that they had been better taught, but this did not seem to be the whole story. Having previously dipped my toe into the practice and theory of educational assessment, as a teacher and curriculum leader, as an A level examiner, and during the course of my MSc in Educational Assessment at Durham University, I wondered what professional and academic literature might reveal about how students perceived and experienced questions in GCSE examinations. ‘What,’ I asked, ‘does the research say?’ A few searches, using Google Scholar and university libraries’ online catalogues, suggested that existent research said very little. Almost no-one, it appeared, had asked the students. Examiners and researchers had had their say, but the voices of students were missing, unheard. To me this seemed paradoxical, even unjust. These 180 teenagers emerged from the examination hall as ‘travellers from an antique land,’ in the words of the poet Shelley. Although public examinations had not been

¹ *Ozymandias*, by Percy Bysshe Shelley, published in *The Examiner of London*, 1818

introduced when Shelley wrote his poem *Ozymandias*² (Oxford and Cambridge local exams began in 1858, with Science and Art exams in 1861, and open competition for the Civil Service from 1855: Barnard, 1961, p. 111), the format of formal written exams has not changed much since his time: almost 200 years later, my students had written in silence and against the clock in response to unseen printed questions set by a distant examiner.

According to the UK government, over half a million teenagers sit GCSE examinations in England, Wales, and Northern Ireland each year; the results determine their futures, and they also dictate prospects for their teachers and their schools.³ Despite this, it appeared – at least to me, as the seed for this present study put forth its first root in my mind – that the education profession as a whole did not have a good grasp of the factors that determined how difficult questions actually were for students, and examiners might not know if their questions operated as they intended. And, if the system was not well understood, how could it be fair; how could it possibly be improved?

GCSE examination questions are used in contexts other than the examinations for which they were created. Since the stakes for these examinations are high, it is not surprising that examination preparation, through answering past examination questions, becomes an increasingly prominent feature of the teaching and learning regime in secondary schools for the two or more years leading up to the GCSE examinations. Teachers and students become very familiar with the style and format of different questions: many teachers refer to ‘a 4-marker’ or ‘a 6-marker’ in their lessons, and students are trained to recognise the differing expectations and techniques that these questions employ in different subjects. For students,

² Shelley’s poem *Ozymandias* is now well known to a generation of school children, through its inclusion in the poetry anthology for AQA’s GCSE English Literature Past and Present Poetry: Power and Conflict (8702/B/2)

³ According to the UK Government, 622,350 16-year-old students sat GCSEs in 2022, an increase of 1.5% on the previous year. The average number of GCSEs per student was 7.78. <https://www.gov.uk/government/publications/infographic-gcse-results-2022/infographics-for-gcse-results-2022-accessible#number-of-gcse-taken-in-2022-by-16-year-olds-in-england>. Accessed 19.08.2023

then, examination questions are an inescapable part of their school life, and an increasingly prominent feature of their learning experience in their later school years. Views of these questions, from the students' desks, might be quite different from examiners' and teachers' views, and the implications of these views for teaching and learning might be far-reaching. If professional educators understood examination questions better, they might be able to improve their effectiveness, enhance teaching and learning and strengthen the validity of the assessment process. If we persist in sending our children to this "foreign land," as the modern equivalent of an ancient initiation rite, we ought at least to learn more about what it is like for them when they get there. As St Benedict of Nursia realised in the 6th century, young people, when asked, often have valuable insights: their voices should be attended to. For me, this is a matter not just of fairness but of social justice: if we do not understand their perspective, the least advantaged students will be the least well prepared for survival in this alien landscape, and so they will be least likely to gain the grades that could set them up for the future.

This page has been left intentionally blank

Journey to a foreign land

In this study, I will take the idea of a journey as an extended metaphor, a linking narrative that threads through the different stages of this study. The metaphor of life as a journey is a common one; Lakoff and Turner explain:

‘Our understanding of life as a journey uses our knowledge about journeys. All journeys involve travellers, paths travelled, places where we start, and places where we have been. Some journeys are purposeful and have destinations that we set out for, while others may involve wandering without any destination in mind, consciously or more likely unconsciously, a correspondence between a traveler and person living life, the road traveled and the ‘course’ of a lifetime, a starting point and a time of birth, and so on’ (1989, pp. 60-61).

Lakoff (1987, p. 275) classifies the journey metaphor as presenting a *source path goal* image schema, involving a starting point, a route and a destination. I might add that the direction or route map, as well as being more or less planned, may conform to a more or less linear or cyclical model. Turner (1998) suggests that the frequent use of the journey metaphor in education has a (probably unconscious) cultural continuity with ancient Greek educational practices. Just as life is a metaphorical journey, so learning is also a journey, and the quest for understanding is an inner journey. In analysing the constructivist role of metaphors in educational literature, Turner notes that the metaphor is a conceptual frame, enclosing a set of metaphorical expressions that are consistent with the framing conceptualisation. In this study, the metaphor of a journey will be used as an organising conceptual frame, providing a context for each stage of the study of students and their concepts of demand and difficulty in examination questions. In this extended metaphor, the students’ experience of examinations is the “foreign land” I attempt to visit. Double inverted commas surround the metaphorical travel terms in this section.

This study has been an exploration of a land that is at once strange and familiar. When students speak about their experience of examination questions, I feel, like Rossetti, that ‘I

have been here before / But when or how I cannot tell⁴.’ These students’ experiences are different from mine, and they inhabit this land of examination questions right now. It was not always so: once I dwelt there; I sat where they sit now, in that same heightened state, at my own desk, my own square yard of judgement; and I breathed the same mixed air of silent expectation, three parts anxiety and one part boredom. But that was long ago; since then, the land has changed, to some degree, and so have I. As teacher and headteacher, I have taken up instead the role of provisioner, store manager of the travel emporium: for more than 25 years, I have sought to supply what others need for their sojourn in examination land (and for their subsequent journeys), but I do not travel there myself. On the one occasion more recently when I once more sat a formal examination, as part of my MSc in Educational Assessment at Durham University in 2011, I was catapulted back to that land, but with my teenage nonchalance in the face of high-stakes assessment now supplanted by middle-aged hyperconsciousness. That single experience apart, I do not visit the country any longer: to understand the land of examination questions as my students experience it, I need them as native dwellers and contemporary travellers to interpret and report back for me.

In terms of this study, I first “get the travel bug” as I stand outside the examination hall and listen to students talking about the different questions they have just grappled with in the exam. This makes me curious about the “foreign land” of the examination questions from which they have just returned. I have written about this above, and in Chapter 1 I give my introduction and an explanation of researcher positioning. I start to “learn the language” of the foreign country by surveying perspectives on demand and difficulty in Chapter 2. There is a metaphorical question about whether students taking examinations are “visitors to” or “inhabitants of” the foreign country; in this study I take the stance that they are temporary residents, since this is a place where they are required to stay for a period. Rather like an arid desert environment, or the vacuum of space, or a dark and high-pressure environment deep

⁴ *Sudden Light*, Dante Gabriel Rossetti, 1853/4

beneath the sea, this is a fascinating but hostile terrain: only those who need to can stay there, and even then only for as long as they must⁵.

I then attempt to “map the known world” of this country by surveying and reviewing the literature in Chapter 3. I find that, although the exterior of this country is well mapped, much of this is from the viewpoint – and for the benefit – of “commercial travellers” (examiners and researchers); few “journeys to the interior” have been made or documented, and few people have discovered what it is really like to live there. This establishes a gap in knowledge that provides a rationale for my research question. Moreover, the “lack of local knowledge” makes an argument for understanding more about the lived experience of those students who are currently dwelling in that country of examination questions. However, in the words of Braun and Clarke (2022, p. 120), and expanding the metaphor, this thesis does not seek to show that I have ‘found an empty cell in the spreadsheet of ultimate truth about the topic’ of examination questions, which my study will fill; rather, I conceptualise the aim of my qualitative analysis as ‘contributing something to a rich tapestry of understanding that we and others are collectively working on, in different places, spaces and times.’

In Chapter 4, I plan my more immersive journey in detail; this is my methods chapter. I choose “the road less travelled” as in Robert Frost’s famous poem, deciding to use reflexive thematic analysis in a thoroughly qualitative research paradigm as my chosen research method, against a research environment in education that increasingly values quantitative study (see Section 4.8). I embark on my first proper trip, “finding, losing, then finding my way again,” encountering some local difficulties with travel restrictions and disease, in the form of COVID-19 and its effects on formal education. This is my online pilot study in GCSE mathematics examination questions, reported in Chapter 5. I evaluate the lessons I have learned from this

⁵ Ethical discussions around our society’s requirement for each generation of children to make this odyssey to such a hostile environment for the ordeal of public examination belong to a different study, and are not considered here.

expedition, acknowledging the limitations imposed by COVID-19 restrictions, and using these insights to design a more substantial return visit to the country the following year.

In Chapter 6, I explore this strange-but-familiar world in more immersive detail, with in-depth encounters, discussion and insights from the people I encounter and with whom I travel. This is the report of my main study of GCSE mathematics examination questions, conducted with almost a hundred Year 11 students in the comprehensive school in North East England where I am currently the headteacher. I report the voices and experiences of the students, as shown in their responses to questionnaires and their discussions in focus groups. I develop themes from the student responses, and relate them to the “maps and guide books” reviewed in Chapter 3.

I arrive back home and “tell the story of my adventures” in the Discussion, Chapter 7. I use what Braun and Clarke (2022) call *thick* analysis and description that tells a rich and interpretative story of students’ lived experiences. Finally, in the Conclusion, Chapter 8, I “reflect on the journey,” the problems I encountered and how I have been changed by it, and present the lessons and insights for “future travellers.” I acknowledge the limitations of my study, and I bring forward recommendations and describe “opportunities for further travel.”

I hope, by using this linking narrative of a journey, to “take the reader with me” through this study, getting a sense of where we have been and where we are going next.

Tone of voice

In this thesis, a mixture of first and third person narrative is employed. It is a core tenet of the reflexive thematic analysis method (Braun and Clarke, 2022) that the researcher is *present* in their research; often it feels appropriate for me to appear as *I*, presenting my findings and my analysis. At other times, a slightly more detached tone is appropriate, where more objectivity is possible and desirable, and I have adopted the third-person or even passive voice. The choice of *person* is therefore intentional, and I hope it is useful rather than confusing for the reader.

In a study that draws so heavily on the words of others, it is important that the reader understands *who* is speaking, at any point. Quotations are always contained within inverted commas (single speech marks), even when they are paragraphed and indented. Double inverted commas are used for coined phrases. The voices of students are central to this study. Therefore, following Robinson's example, 'throughout, quotations from children themselves are given prominence by placing them *in italics*' (2014, p. 1, emphasis in original).

This page has been left intentionally blank

Chapter 1: Introduction – The idea of a journey

In this chapter I introduce the “foreign land” of examination questions and give my brief autobiography as headteacher-researcher and narrator of the story of the journey.

Within the educational systems of the United Kingdom (England and Wales, Scotland, and Northern Ireland), examiners have traditionally influenced the work of students by determining the curriculum content (the “specification”) for examinations. Fautley points out that ‘what is valued tends to be what is assessed’ (2015, p. 513), so examiners contribute to a sense of what is valued within curricula. Although schools are expected to offer a ‘broad and balanced curriculum’ in students’ earlier years,⁶ the curriculum narrows as the examination years approach in secondary schools, and what is to be assessed directly influences what is taught. The influence of examiners is sometimes referred to as ‘backwash’ or ‘washback’ (Cheng and Curtis, 2004, p. 3; Taylor, 2005, p. 154) and can, at its most extreme, be described as ‘teaching to the test’ (Posner, 2004, p. 749). British secondary school students who are aged 16 sit examinations that are marked by examiners and lead, in England, Wales, and Northern Ireland, to the award of the General Certificate of Secondary Education⁷ (hereafter GCSE).

The UK examination system also provides material for research. Examination questions, results, and trends are studied by researchers using quantitative and qualitative methods. To complete the triangular relationship, researchers, through their influence on teachers and teacher training, may also affect the ways in which students are taught. This relationship and dynamic is considered at greater length in Chapter 8.

⁶ See, for example, Department for Education guidance, 2021: <https://www.gov.uk/government/publications/teaching-a-broad-and-balanced-curriculum-for-education-recovery> accessed 12.05.2022.

⁷ Children in Scotland take a different suite of qualifications. Most children in Scotland take National 4 or National 5 examinations, at the age of 15 rather than 16: <https://www.theschoolrun.com/overview-scottish-education-system> accessed 12.05.2022.

This study advances existent academic knowledge with regard to how secondary school students understand the ways in which GCSE examination questions work in the core subject of mathematics. I will argue that the interplay between concepts of “demand” and “difficulty”, is currently not understood sufficiently well by students or by the “educational community” (for the purposes of this study, defined as being made up of teachers, examiners, and researchers). Through a review of literature, I will demonstrate that such understanding of “demand” and “difficulty” as does exist is unbalanced: it comes from the perspective of examiners and researchers, and not from students. Existent understanding is also partial, as evidenced (in Chapter 2) by the lack of shared and commonly accepted meanings of even the core vocabulary of assessment. The implications of this lack of shared understanding are that the GCSE examination system does not work as well as it could, and that students are at a distinct disadvantage in approaching their GCSE examinations because they do not fully understand what the examiners are seeking (Wood, 2007). Cumulatively and individually, the results of GCSE examinations in mathematics as well as in other subjects, determine the ‘life chances of individual test takers’ (Taylor, 2005, p. 154), and they have implications for the status and success of teachers and their schools (Standish and Perks, 2021; Kellaghan and Greaney, 2019). It is therefore critically important that the questions asked in GCSE mathematics examinations “perform” as expected, so that students can both give answers that accurately demonstrate their knowledge and expertise, and secure examination grades that give appropriate credit to their knowledge and expertise.

This thesis undertakes a critical assessment of the understanding that examiners, researchers, and students have of the concepts of “demand” and “difficulty.” In so doing, it reveals the weaknesses and failings of existent shared understanding, and evaluates what might be meant when the two concepts of demand and difficulty are discussed. In consequently making recommendations for improvements to the system of examinations, and advancing suggested avenues of further study, my study suggests ways by which the relationship that exists between examiners, researchers, and students can be rebalanced.

1.1 Context: the GCSE examination system in England, Wales, and Northern Ireland

The system of public examinations leading up to the award of the GCSE in England, Wales, and Northern Ireland is very familiar to those who have regular contact with secondary school education, such as students, parents, teachers at schools and universities, examiners, and researchers. The GCSE examination system is so familiar that it is, on the whole, unquestioned, except when something unusual happens such as a dramatic change in the grades awarded from one year to another, or the cancellation of examinations in the time of the COVID-19 pandemic. Such events bring the examination system into the collective consciousness, after which it subsides once more. Given this, the first task that this study undertook was to probe into the examination system to reveal its hidden workings so as to make 'strange what had appeared familiar' (Elliott, 1994, p. 424). A simplified model of an examination system was constructed, and the workings of the existent system were compared with this simplified model. Relevant literature was reviewed: to evaluate how examiners and academics understand concepts of demand and difficulty operating within the examination system; to evaluate the interaction between learning theories and the way that examination questions work in practice; to understand how validity theory applies to examinations; and to recognise ways in which students' views about examination questions can be acknowledged. Following this, a practical study was undertaken in two parts, involving students evaluating examination questions in mathematics. Finally, the thesis presents the conclusions generated by this study and, from the new knowledge and understanding generated, makes recommendations to improve both the examination system and methods of teaching and learning that prepare students for examinations.

1.2 Socio-political aspects of the current high-stakes assessment system

The current system of closed-book examinations is both familiar to many people in the United Kingdom and, to a large extent, unquestioned. Given the prevalence of the closed-book invigilated examination within the Anglophone western education system, it would be easy

(but erroneous) to assume that “high-stakes” examinations are an evident good. Their continued place in a “knowledge economy,” however, is not without controversy. It is not the purpose of this study to balance the claims of those who stress the importance of acquiring biologically secondary knowledge against those who assert that we ought to equip our children with skills to tackle the unknown needs of the future. If ‘schools were invented specifically to teach biologically secondary, culturally specific knowledge that students are not motivated to learn independently’ (Didau, 2019, p. 54), then closed-book, invigilated, final examinations are a secure way to test whether that knowledge has been learned. However, there may be other, more adaptive, ways of doing this. Biggs (1993) and Ramsden (1992) proposed that knowledge is of less significance than “meaning,” and that meaning is not imposed or transmitted by direct instruction, but is constructed by the students’ learning activities. Herrington and Standen (2000) commended a constructivist paradigm for assessment, emphasising the role of the metacognitive process in knowledge construction, rather than focusing on the material that has (or has not) been learned. Williams goes further and argues that,

‘In the information age the closed book, invigilated final examination has become an anachronism. Most significantly, it is an assessment instrument that does not assess deep conceptual understanding and process skills. Indeed, the anecdotal evidence one often hears from students is that ‘cramming’ the night before amounts to ‘data dumping’ on the day, with little knowledge retention thereafter. The defence of the traditionalists is that we have to have invigilated final examinations or students will cheat’ (2006, p. 107).

If, taking Churchill’s 1947 much-cited line about democracy, closed-book invigilated final examinations are the worst system of assessment – except for all the others – what are the alternatives? Rapke reported on a Canadian study in mathematics, where students assisted their teacher in developing the final closed-book examination that they then sat. Claims were made that the students experienced deep approaches to learning as they ‘worked as partners in learning, teaching and assessment during the process’ (Rapke, 2016, p. 27). A bright new dawn of assessment reform was hailed by some optimistic educationalists as one potential benefit of COVID-19 (Akulwar-Tajane *et al.*, 2021; Meeran and Davids, 2022), proposing open-

book and/or online examinations. This was partly for pragmatic reasons, but also as an opportunity to reform examinations. The approach of examination boards in England, post-COVID, was to thin out some examined content but otherwise to retain the structure and format of examinations. In the aftermath of a globally disruptive event such as the COVID-19 pandemic, with its dire economic and social consequences, few political leaders appear to have had either the appetite or the available finance for radical reform of structures such as schooling and examinations.

Following an interest in experiential learning (Bruner, 1967), learning styles (Coffield *et al.*, 2004) and problem-based learning, the increase in attention given to cognitive science in initial teacher education (see, for example, Kirschner and Hendrick, 2020) has meant that a pendulum in educational thinking has swung in favour of knowledge acquisition and helping educators better understand ways to help students learn effectively. This has tended to reinforce the position of formal public examinations in our education system, as a known system to test, however imperfectly, the extent to which that knowledge has been acquired.

For schools leaders, public examinations are expensive⁸ and they impact on mental health in ways that are well known to students, school leaders, parents and the government⁹; but they also provide some strong positives, including offering students qualifications that are recognised and regarded as robust by employers and parents. Public examination results provide accountability for schools, enabling parents, inspectors and national leaders to compare the performance of individual schools. Departing from GCSE and other recognised

⁸ As headteacher of a medium-sized comprehensive school, I allocate a budget of around £150,000 each year to examination fees for GCSE, A level and vocational qualifications, plus staff costs for an examinations officer and a team of invigilators

⁹ <https://www.theguardian.com/education/2021/sep/06/exam-system-in-england-needs-an-overhaul-says-schools-leader>; <https://ofqual.blog.gov.uk/2019/03/15/what-can-schools-do-about-examination-and-test-anxiety/> both accessed 04.07.2023

qualifications assessed through examinations would carry considerable risk for schools, and few if any are able to contemplate this¹⁰.

1.3 How examination questions should work in a simplified system

When thinking about examination questions, and examinations as a whole, it would be easy to entertain a naïve or simplified view of an examination system. Such an examination system might appear, to an observer, to be akin to a factory process: a machine, almost, with inputs, a process, and outputs. In this simplified model, examination questions would be the ‘inputs’; the ‘process’ would be the sitting of examinations and their marking and standardisation; and the ‘outputs’ would be the students’ grades. These output grades (whether in the idealised or actual examination model) stand as proxies for a comprehensive measure of students’ cognitive abilities. This simplified model (named here, not without irony, as the *Utopia Ltd Examinations System*), is visualised in Figure 1.

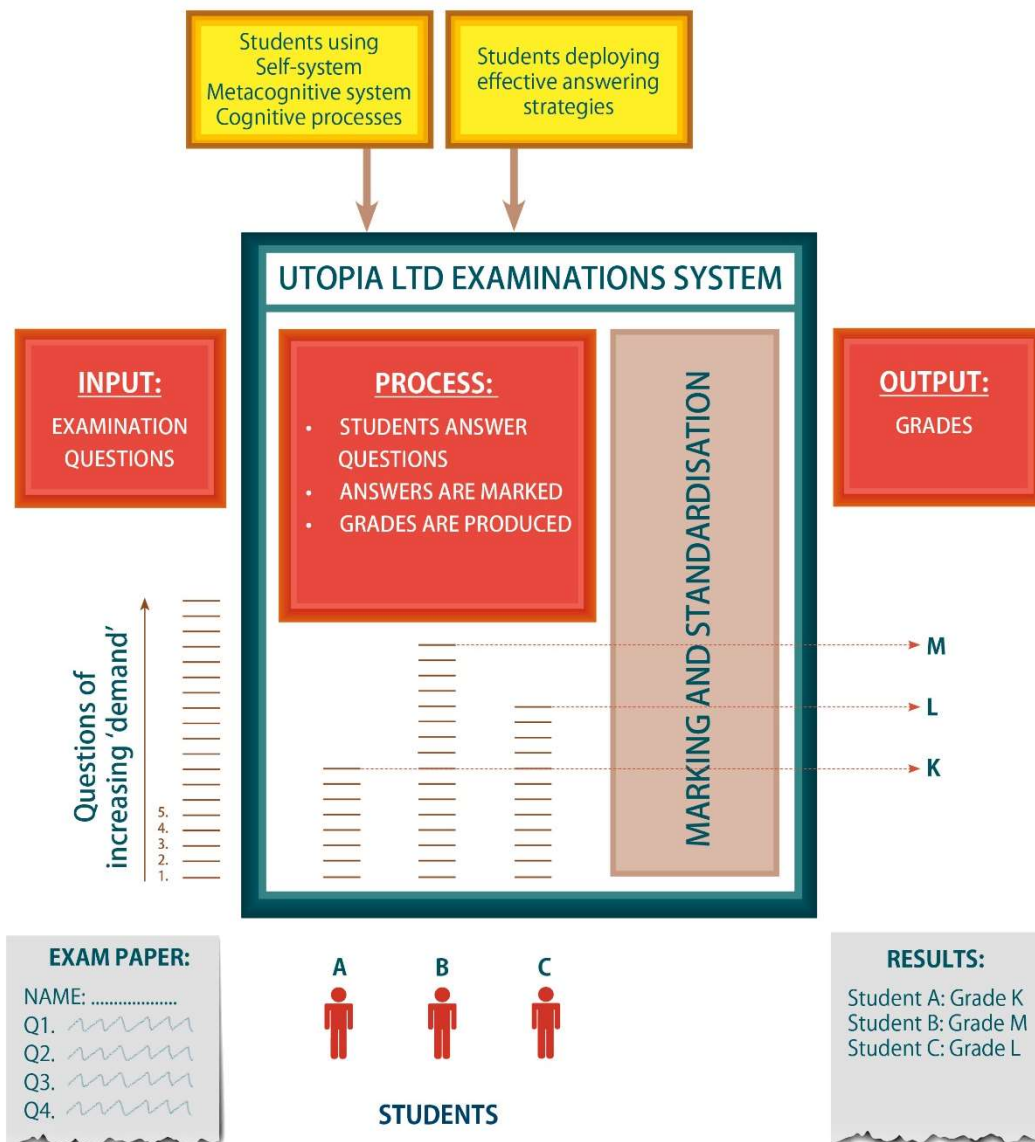
Within this naïve and simplified model, questions are arranged within examination papers in order of increasing demand. The demands have been planned by the examiners: they can predict how difficult they expect students will find each question. These questions perfectly sample the domain of the syllabus (or specification) of the given subject and therefore ensure strong content validity. There are questions at every level of demand to match against students of all levels of ability and expertise. Due to the fact that there is a perfect match between the examiners’ intended demands of the given examination paper and the students’

¹⁰ The International Baccalaureate suite of qualifications offers an interesting comparison. 38 state schools in England offer the IB Career-related Programme, and 22 offer the IB Diploma Programme, both of which are externally validated and internationally recognised. However, although 16 English state schools offer the IB Middle Years Programme (IBYP), which is an unvalidated curriculum programme available for students in years 7 to 11, none of these schools offers the IBYP beyond Year 9: all these schools transfer students to GCSE and vocational qualifications in Year 10. Numbers of schools from <https://whichschooladvisor.com> accessed 04.07.2023

experienced difficulties, valid inferences will be able to be made about each student's ability from the grade they obtain.

Students, who, in this idealised system, are well-prepared and who approach the examination in a calm state, bring differing levels of expertise and prior learning. Some, such as Student A, are able to answer the less demanding questions correctly, but do not progress beyond these. Others, such as Student C, progress further, answering all the less demanding questions correctly, as well as some of the more demanding questions. A few, such as Student B, manage to answer substantially all of the questions correctly, including some of the most demanding ones.

Figure 1 - A Simplified Examination System



Source: Author's own (graphic design by Jo Murray)

Many different parts of this process can be studied and, indeed, have been: the creation of examination questions; the preparation of students; marking and standardisation; the production of grades according to national criteria, and so on. The focus of this study is what happens within the left-hand side of the examination system, the 'process' box; that is, the interplay between examination questions and the students who answer them.

Black and Wiliam applied a “systems engineering” analogy to teaching and learning, and suggest that,

‘Present policy seems to treat the classroom as a *black box*. Certain *inputs* from the outside are fed in or make demands – pupils, teachers, other resources, management rules and requirements, parental anxieties, tests with pressures to score highly, and so on. Some *outputs* follow, hopefully pupils who are more knowledgeable and competent, better test results [*etc.*] ... but what is happening inside? How can anyone be sure that a particular set of new inputs will produce better outputs if we don’t at least study what happens inside?’ (1998, p. 1; emphasis in original)

In a similar way to Black’s and Wiliam’s enquiry into the ‘black box’ of the classroom, this study focuses on what goes on inside the examination process box, and specifically about what goes on in the mind of the student when answering examination questions. Just as Black and Wiliam wished to disassemble and examine what went on in the ‘black box’ in order to improve their understanding of how learning happens, in this study I want to travel to the remote territory of examination questions, to hear from students about what happens there.

When a student addresses an examination question, whether in the context of an actual examination or within the classroom as part of a process of examination preparation, their teachers will hope that “hard thinking” is activated (Coe, 2015, p. 13).¹¹ Examination questions, whether used formatively in a classroom or summatively in a GCSE examination, are used to activate cognitive processes. Given that it is not feasible to travel, whether figuratively or literally, into the minds of students as they take their actual GCSE examinations, asking students about their experiences of examination questions provides a lens through which an understanding of their thinking can be observed. In adopting this approach, this study makes the axiomatic assumption that answering examination questions involves students (both individually and collectively) thinking in ways that are very similar if not identical to those

¹¹ Coe: ‘Learning happens when people have to think hard,’ presentation to IB World Regional Conference, Den Haag, Netherlands, 31.10.2015. <https://www.ibo.org/globalassets/events/aem/conferences/2015/robert-coe.pdf> accessed 07.05.2022. This presentation was linked to the publication of the Sutton Trust/Durham University CEM report ‘What makes great teaching’ by Coe *et al.*, (2014).

deployed in the learning process, but with the proviso that the context of a high-stakes examination may add additional stresses.¹²

When answering examination questions, whether in actual examinations or practising within lessons, or revising, students engage their cognitive thought processes. A thorough understanding of these cognitive processes is, therefore, vital to this study's discussion of how examination questions actually work. The cognitive processes that operate within the minds of students are explored and, in the survey of published literature presented in Chapter 3, models and theories that classify and explain these processes are critically evaluated.

This study contends that not enough is known about students' perspectives of their own cognitive processes. As will be discussed in Chapter 3, there has been very little research into either the strategies that students deploy to answer questions, or the understanding that they have about the factors that make questions more or less demanding. It is into this gap within existent understanding that this study makes a contribution of new knowledge, contextualising it within existing knowledge and theory.

The assumption that thinking about examination questions uses the same cognitive processes as those used in learning is reasonable, given that public examinations can be seen, within a compulsory education context, as the last summative assessment exercise in a long process of teaching and learning (Black *et al.*, 2003). In lessons, there is at first more teaching and less performative activity. Later, students practise and improve their understanding by answering questions and receive feedback.¹³ Teachers regularly use examination questions in this context, up to two or three years before the GCSE examinations themselves. The feedback given on practice questions, although summative in nature (that is, the feedback is reflective,

¹² The link between cognition and examinations is explored in more depth within Chapter 3, the Literature Review.

¹³ This approach to teacher modelling is summed up in an "I do – we do – you do" approach, as explained by Evidence Based Teaching: <https://www.evidencebasedteaching.org.au/the-i-do-we-do-you-do-model-explained/> accessed 02.06.2022 (see also section 3.1.2 later in this study).

backward-looking), acts in a formative manner (Black *et al.*, 2003)¹⁴: it should enable students to learn from what they have just done and move on to be more effective. Examinations, as purely summative assessments on which no feedback is given to students other than a grade, complete the transition from teacher input to student demonstration. Since the role of examination questions is central within a wide range of teaching and learning activities, a better understanding of how examination questions work, as attempted in this study, is therefore vital.

Within the simplified system of examination exemplified within this thesis by the *Utopia Ltd Examinations System*, a simplified marking and standardisation process would take place once students answered the examination questions of increasing demand. The answers given by the students would be marked, moderated to ensure reliability and accuracy, and then standardised.

Standardised marks would then, within this simple system, be converted into grades. These grades would be reported as results. Graded results are the outputs of the examination system. In the simplified examination system presented in Figure 1, grades have been given as letters, with letters that come later in the alphabet equating to higher grades.¹⁵ Utopian student A, who managed to answer only a few questions correctly, would gain a relatively low grade; (Grade K within Figure 1), whereas Student C who answered the less demanding questions correctly and went further, would gain Grade L. Student B, who was evidently the most expert, because they managed to answer correctly not only the less demanding questions but also some of the most demanding questions, would gain the highest grade; Grade M.

¹⁴ 'The aim of summative assessment is generally to report on students' level of learning at a particular time, rather than to impact on ongoing learning, as in the case of formative assessment' (Dolin, *et al.*, 2018, p. 61).

¹⁵ The grading system given here is an imaginary conflation of lettered grades, in use in GCSE until 2019-20, where A* was the highest grade and G the lowest, and the more recent numbered grades, which work in the opposite direction, where grade 9 is the highest grade and 1 the lowest.

In this simplified examination system, it is straightforward to interpret the grades K, L and M. They represent readings of the expertise and knowledge of the three different students in the given subject under examination conditions at the time. These interpretations have strong validity: an argument can be made that validates how the grades were obtained, how they relate to the knowledge domain, and what they represent. There is, in the simplified system, an evidential link between the input questions and the output grades. It follows that users of the grades can be confident in the validity of the interpretations they make based upon them. It may be possible to go further, and to infer that the achievement of a high grade in the simplified examination may predict future success in the same subject (or other subjects), or that it may be an indicator of other expertise. These inferences carry validity that is less strong, because they rest on assumptions rather than evidence, but an argument can still be made for them. This validity argument is reviewed in Section 3.3.

The Utopian system can be seen as “pure”, free from sources of error; a pure “signal”, free from “noise”. Unfortunately for everyone concerned, the *Utopia Ltd Examinations System* does not exist, because human beings and the systems they create are neither consistent nor perfectly predictable. As evidenced by the large body of published research into examinations and formal assessments, the system within the UK for GCSE examinations is only an approximation to an idealised system (see, for instance, Dhillon and Richardson, 2003; Broadfoot *et al.*, 2012; and Pollitt, 2012). It contains sources of error, or noise; there is grit in the machine. Sources of unwanted error include: questions where the demand is unknown and/or unpredictable; questions that appear to distract or mislead students through their presentation or wording which may, in turn, contribute to unexpected difficulty; students who are affected by performance issues such as anxiety or examination stress; and marking and standardisation systems that may be unreliable or biased. These imperfections, and their implications, are discussed in Chapter 7.

The result of such factors is the existence of an imperfect system examination system. This, in turn, damages the system's validity. If the system does not perform perfectly, then there cannot be absolute certainty that the grades that form its output reliably indicate the expertise of the students (Wood, 1991; Stobart, 2008). The grades may not, therefore, be valid indicators of either expertise or future success. Nevertheless, these output grades are vitally important: the stakes are high for students, their teachers and for schools, but there is a danger of 'over-simplistic interpretations, which may claim more than they can justify' (Stobart, 2009, p. 171).¹⁶ Given this, it is important that a detailed understanding is gained into how the actual examination system works, so that it can be made to work better for everyone in the future. In furthering existent academic knowledge, this study increases the understanding of researchers, examiners, and teachers in the hope that the improved understanding may lead to improvements in the examination system as well as in teaching and learning. In offering an interpreted version of students' experiences, I aim to contribute to a greater understanding of their lived experience as it relates to examinations in schools. By informing teachers, students, and examiners, the ultimate aim of this study is, therefore, to help strengthen the validity of the GCSE examination system, as used in England, Wales and Northern Ireland, and to bring into the collective conscious the experience of students who are at the heart of this system.

The terms "demand" and "difficulty" have already been introduced; definitions of these concepts are cited and discussed in detail in Chapter 2. These two concepts are central to the examinations system. Demand ought to be in the hands of skilled examiners, who choose the level at which to pitch each question. It should be possible to create a demanding question on an easy topic, therefore, and *vice versa*. Looking from the other end of the process, difficulty can be measured as an inverse function of how many students correctly answer a question. It

¹⁶ Stobart was referring specifically to National Curriculum tests, but his comments also apply to GCSE.

is an inverse relationship because a difficult question will be correctly answered by fewer students.

1.4 Rationale for this study

Examiners, researchers, and teachers may assume that they know how examination questions “work” – specifically, how question characteristics relating to demand translate into difficulty for students – and an important part of the discussion pertaining to definitions in Chapter 2 is the understanding that examiners and researchers appear to have. Nevertheless, in the absence of empirical evidence, it is far from certain that questions do in fact operate in the way that examiners intend. To find out more about the real world operation of examination questions, it is necessary to engage with students themselves. Student voice, therefore, lies at the heart of this study’s quest for greater understanding.

In section 1.3, a simplified examination system model was presented. Within it, examination questions play a central role. On the surface level, ‘questions in an examination are the problems which are set in order to test your knowledge or ability’ (Collins Dictionary, 2022).¹⁷ By probing a little deeper it can be argued that the fundamental purpose of an examination question is to activate students to think, so that they may produce outputs that can be assessed and reported in a way that contributes to a fair judgement of their individual cognitive ability. This basic operation of examination questions holds true regardless of context. A student may encounter a question within a public examination, within a mock (practice) examination, within a lesson – where it may be used by the teacher to check understanding or to prepare students for a future examination – or within a programme of revision undertaken by the student themselves. There are different question types: some will require a student simply to recall knowledge, while, at the more demanding end of the scale,

¹⁷ Collins Dictionary, online, accessed 12.05.2022.

other questions will require students to analyse and interpret information presented in the question, to synthesize this new information with previously learned and recalled knowledge, and to create an answer that displays a deeper understanding. Different question types are used within this study. All of them, however, operate in a similar manner: each individual question poses one or more demands and, as these are encountered by students, demands translate into different levels of difficulty.

With regard to investigating students' understanding, it is not possible, even with the most advanced medical technology, to look into their minds and understand their cognitive processes as they encounter questions in an actual examination. There would be strong ethical objections to doing this in any case, in the context of high-stakes examinations that may determine students' life chances. Moreover, any such observation might well alter the behaviour of the students (both individually and collectively), according to the Hawthorne effect¹⁸. Instead, a practicable approach is to ask students direct questions about their views of the demands and difficulty of individual examination questions, and to do so outside the context of the (given) examination. This is the approach taken in this study. The relative strengths and drawbacks of this approach are noted in Chapter 4.

Individual students are likely to display a range of different responses to particular examination questions, and they may also articulate their understanding in different ways. This study captures a wide range of responses and understanding. This was achieved by operating across a reasonably large sample size, and drawing out common and contrasting threads of understanding. Recruitment methods and sample sizes are also discussed in Chapter 4.

Due to the fact that GCSE examinations represent a final, summative assessment of students' understanding and expertise, it might be expected that the focus of this study would be educational assessment. However, that is not the purpose of this study. Rather, its rationale is

¹⁸ 'The Hawthorne effect is when there is a change in the subject's normal behaviour, attributed to the knowledge their behaviour is being watched or studied' (Oswald, *et al.*, 2014, p. 53).

to understand more about how students respond to individual examination questions, whether they encounter them in a lesson, a revision session, or an examination. From my perspective, as a teacher as well as a researcher, I argue that the function of an examination question, regardless of context, is to activate thinking in students so that their individual and collective cognitive ability can be measured. Given this, this study therefore locates its primary focus in the field of teaching and learning rather than educational assessment.

1.5 Researcher positioning

Braun and Clarke assert that ‘the researcher’s positioning inevitably shapes their research and engagement with data’ (2022, p. 14). In qualitative research, this subjectivity is viewed as something valuable rather than problematic, as ‘the researcher becomes the instrument for analysis’ (Nowell *et al.*, 2017, p. 2). Nonetheless, ‘owning one’s perspective’ is necessary (Elliott *et al.*, 1999, p. 220): the researcher is required not just to acknowledge their own subjectivity, but also to interrogate it within an ongoing process of “reflexivity” (Braun and Clarke, 2022, pp. 12-22). Several authors on qualitative analysis, including Braun and Clarke, suggest the keeping of a “reflexive journal”, in order that the researcher can

‘Recognise and take responsibility for one’s own situatedness within the research, and the effect that it may have on the setting and people being studied, questions asked, data collected and its interpretation’ (Berger, 2015, p. 220).

Having been introduced to this practice by one of my later supervisors, I began to keep a reflexive journal during the latter parts of my data gathering and thematic analysis processes, in order to contextualise and reflect on my own philosophical position, theoretical assumptions, ideological and political commitments, social identities, and personal assumptions.

Reflecting on my own situation, I recognise that, as a heterosexual white male from the British middle classes, educated at a selective school, culturally and philosophically Christian, non-disabled, non-migrant, and now in a well-rewarded position of responsibility within an established profession (as a headteacher of a secondary school), I am in a position of

considerable social privilege. In terms of social marginality, however, I consider myself also to be somewhat marginalised in several different ways. Having been brought up in a religiously non-conformist and politically left-leaning household, I came to question most social norms and political received wisdom from an early age. Education, marriage and my career took me to many different regions of England, including the midlands, south west, north east, home counties, north west and south east. The effect of so many moves (15 employers, 13 houses) is marginalising, denying a sense of belonging to any given locality. Although I work daily with children, both as a researcher and as a teacher-headteacher, I am myself childless, by circumstance rather than by choice. My own personality tends towards introversion. It is therefore not surprising that, as an individual and therefore also as a researcher, I am accustomed to the feeling of being an outsider in most situations.

These elements of privilege within a context of slight marginalisation provide a distancing perspective – an ‘intimate distance’ (Pile, 2010, p. 483) – that may be of benefit to a researcher. The start of this research project coincided with my obtaining my first position as headteacher of a secondary school, meaning that I have occupied the role of headteacher-researcher, in two contrasting school situations, for the entirety of this project. This is not a *dual* role, but rather an *integrated* role: the two aspects – headteacher / researcher – are not separable, and they exist in a symbiotic relationship, each sustaining the other. For a teacher within a school where the research is situated, there is a privileged position of access to the research subjects (students) and the advantage of familiarity with their learning environment. This brings the researcher and their research subjects closer together. The position of authority that the headteacher holds, however, by dint of their office, introduces a distancing effect into this relationship. Feminist and indigenous iterations of reflexivity emphasise that power is part of knowledge production (Ramazanoglu and Holland, 2002; Russell-Mundine, 2012). Insights from these perspectives caused me to interrogate my own position and the relationship between researcher and research “subjects” (a power-laden word) in the current study, and to consider the politics inherent in both the research process and the knowledge

that is produced. One of my students told me, after he had completed the questionnaire, that he had 'tried harder, and filled in all the sections, because I knew it was you who would be reading it.' I interpret this to mean that he might not have put in so much effort for an online form or if he did not know the researcher. These reflections, as well as a consideration of the power relationships implicit in the examination system, are discussed further in Chapters 7 and 8.

The position of headteacher-researcher may be an unusual one, but it can be understood as an extension of the more common teacher-researcher role. Baumann and Duffy (2001) note that there is a long and rich history of teacher action research, as well as a more recent resurgence of interest in teacher research. According to their analysis of 34 teacher research studies, the teacher researcher typically

'Identifies a persistent teaching problem or question and decides to initiate a classroom enquiry. This teacher reads theoretical and applied educational literature, including other teacher-research reports, and decides to work collaboratively with a colleague. Using primarily practical, efficient, qualitative methods, recommended by other teacher researchers, with perhaps a quantitative tool added in, the researcher initiates a study. The teacher learns from and along with students while engaging in the investigation, and she or he finds that the research questions have been altered somewhat throughout the course of the study. The researcher may struggle to balance the dual role of teacher and researcher or may feel uneasy with the innovations that are explored. The teacher researcher decides to share the research story publicly and writes it for publication, using a narrative style that includes figurative language and verbal and visual illustrations' (Baumann and Duffy, 2001, p. 611).

The present study has many similarities with the approach described here, as well as some differences. In common with Baumann's and Duffy's generalised teacher-researcher approach, as headteacher-researcher I here identified a question (albeit a more systematic one) and decided to initiate a classroom-based enquiry. I read a range of relevant theoretical and applied literature, and I initiated the study using primarily qualitative methods with also a quantitative tool (see Chapter 4: Methods). The research question was altered and adapted during the course of the study. As researcher, I learned from and along with the students, and I

decided to share the research story publicly. Verbal and visual illustrations were used in the write-up.

There were some differences from Baumann's and Duffy's generalised approach, however. Unlike many teacher-research projects, the nature of this enquiry did not lend itself to an action-research approach. Opportunities for researcher collaboration in the present study were rare. As headteacher-researcher, I shared – and, through my influence on school policy and the professional development of teachers, also shaped – the educational environment of the classes I visited, but I remained an outsider to the students' social and class groupings. Research findings were created through an inductive thematic analysis approach, where the headteacher-researcher acted as an intermediary and interpreter, familiar with both the student and the research context, without being fully a member of either community. Following the tenets of 'Big Q' qualitative research (Braun and Clarke, 2021a) my role was not to be invisibly objective, but to be visibly present in the process, making plain the hidden assumptions and processes of students as they approached and "solved" examination questions. My subjectivity in this process of thematic analysis was therefore a strength, rather than a weakness: my headteacher-researcher's immersion in the context of the students enabled me to present the "story" of the students' struggle with examination questions, so that the findings of themes could be seen as 'data with soul' (Nadar, 2014, p. 18). I can then offer this interpretation to others who are either further "outside" the classroom situation – educationalists, researchers, and examiners – or further "inside" – teachers, school leaders, parents and, of course, the students themselves.

It is entirely possible that another researcher might construct a different set of meanings and stories from this data set, with different validity claims. *This* headteacher-researcher, working with *these* students in *this* context, constructed *this set* of meanings and insights and, as contributions to the rich tapestry of understanding about the role and function of assessment within teaching and learning, they bring new knowledge into the field.

1.6 Research aim and question

As a headteacher, I set out wishing to understand more about my students' experiences and perspectives on the examination questions they encountered. My study is shaped to some extent by what I bring to it as well as the stories and meaning I develop within it. This approach is entirely in keeping with the 'Big Q' qualitative orientation of my study: analysis is viewed as a process of meaning-making rather than truth-seeking or discovery (Braun and Clarke, 2021b).

As Nadar (2014) argues, when telling the stories of those whose views have been under-represented in the forum of academic discourse, subjectivity is a strength, rather than a weakness. Nonetheless, great caution and self-awareness needs to be taken by the qualitative researcher, since otherwise 'the analysis and findings may say more about the researcher than about the data' (Cohen *et al.*, 2018, p. 666). On an inductive-deductive spectrum, the research presented here leans more to the inductive end: in investigating students' experiences, I did not know what I would discover, and the analytic codes and subsequent themes were developed from the students' responses. In order to accommodate and encompass a breadth of experience and meaning, the research question needed to be broad enough.

Two concepts therefore drive this study: the students' *experience* of examination questions, and my *understanding* of this experience. At the heart of this study is the research question:

- **How do students experience and comprehend demand and difficulty in GCSE mathematics examination questions?**

The choice of verbs in this research question is deliberate. "Experience" is defined (Meriam-Webster, online) as 'to come to a knowledge of (something) by living through it.' A central aim of this study is to capture, distil and develop the *lived experience*¹⁹ of students, and to offer this meaning as a contribution to the rich tapestry of knowledge about assessment. The word

¹⁹ Meriam-Webster (online) also offers the following highly relevant synonyms for "experience": endure; undergo; witness; taste; see; feel; suffer; encounter – all these verbs mediate individuals' lived experience; many of these verbs have connotations of living through an ordeal or challenge, which also seems appropriate for students' experiences of high-stakes examinations.

“comprehend” has been preferred to the more commonplace and inexact word “understand²⁰,” bringing connotations of getting to the bottom of something, unravelling and deciphering its meaning and importance.

1.7 Impact of the COVID-19 pandemic on this study

The development and progress of this study and, as a consequence, its final shape, was heavily affected by the COVID-19 pandemic. It had originally been my intention to carry out a large-scale survey of students in their schools, and permission for this had been sought and granted from headteachers in the Gateshead local authority area of North East England in 2019.

However, schools in England were closed to all but a small number of vulnerable students, from 20 March to 9 June 2020 and again from 6 January to 8 March 2021, as part of a programme of national lockdowns to curb the spread of coronavirus.²¹ Teaching was conducted remotely during these lockdowns, and even in the months that followed their reopening, schools were closed to all but essential visitors. My plans to visit the schools in person, to present and explain my project, to answer questions, and to supervise the completion of survey questionnaires, therefore had to be abandoned. This was a major setback to the research project.

Given the changed working realities of the pandemic, it was necessary to change the medium of the research tool for the pilot study, with students being engaged through online questionnaires. The limitations of this online method are discussed in Chapter 4. Online questionnaires were emailed to heads of department in the participating schools, and distributed by them to their students. The change of delivery mechanism also had an impact on the design of the research instrument; examination questions needed to be relatively short

²⁰ Synonyms of “comprehend” are given (Oxford Languages, online) as grasp; take in; apprehend; follow; make sense of. Meriam-Webster (online) gives synonyms of decipher; grasp; recognize; appreciate.

²¹ Information from HM Government, UK. Infographic of lockdown timelines at <https://www.instituteforgovernment.org.uk/sites/default/files/timeline-lockdown-web.pdf>

and could not contain graphs or images. Students who responded were self-selecting, rather than being in whole class cohorts. Although the survey design differed from what had originally been intended, it was still possible to secure interesting and coherent responses, and these responses then helped to shape some of the questions asked in the main part of the study.

More significantly for students, GCSE examinations were cancelled for both 2020 and 2021 summer sessions, with systems of grading based on teacher assessments being used instead.²² Although the system of marking and grading changed in 2020 and 2021, however, the role and function of GCSE examination questions did not change substantially. Most students in most schools in England and Wales had taken practice (mock) examinations in the periods before schools had been locked down and they had done so using past paper GCSE questions. Other students had answered GCSE questions in lessons, under controlled conditions. Answers to GCSE questions, created either under examination conditions or as part of classwork, formed the basis for these teacher-assessed grades. Even during periods of national lockdown, then, examination questions remained an important feature of students' educational lives. As such, they therefore remained a significant topic for research.

1.8 Structure of Thesis

Having outlined the aims, rationale and motivation of this study, the rest of this thesis is structured as follows. In the prefatory section on a Journey to a foreign land, I have already outlined the extended travel metaphor that creatively and conceptually frames this study; in this section, the description is more prosaic. Chapter 2 addresses the definitions of "demand" and "difficulty" that are used throughout this work. Thereafter, Chapter 3 explores a range of relevant literature. Because this study explores an area of practice at the intersection of

²² The systems for calculating and awarding GCSE grades in 2020 and 2021 were based on teacher assessments, with Centre Assessed Grades (CAGs) in 2020 and Teacher Assessed Grades (TAGs) in 2021.

different academic fields, the scope of this study's literature review is extensive, although by no means exhaustive. It incorporates, amongst other elements, cognitive load theory, taxonomies of learning, validity considerations and literature relating to student voice.

Chapter 4 discusses the methods used in the pilot and main studies, and also comments on philosophical positions and ethical considerations. Two empirical studies were undertaken as part of this study, both using examination questions from GCSE mathematics papers as their starting point. The first, a pilot study, was undertaken while schools were still under restrictions as a result of the COVID-19 pandemic. These restrictions impacted on the survey method and the quantity and quality of the data gathered. Results and findings from the pilot study are presented in Chapter 5. Lessons learned from the pilot study are evaluated at the end of the chapter, and brought into the design of the main study. The main study was undertaken in the autumn term of 2022, after the lifting of COVID-19 restrictions. The role of the headteacher-researcher was developed further for this study, and focus groups as well as questionnaires were used. The results of the main study are reported in Chapter 6 in rich detail, and the themes of students' responses are developed. The results of the two empirical studies are subsequently summarised, discussed and interpreted in Chapter 7, with a particular emphasis on the implications of the study's findings for students, teachers and examiners, and for the improvement of the examination system.

Having examined and evaluated the views and perceptions of students, Chapter 8 brings the thesis to a Conclusion. In this section, I reflect on the study process, problems encountered and insights gained for future qualitative studies in this area. I summarise the new knowledge created by the study, reflect on meaning developed through the study, and offer some recommendations. Limitations of the study are considered, and opportunities for future study are outlined.

This page has been left intentionally blank

Chapter 2: The concepts of “Demand” and “Difficulty” – Learning the language

In this chapter I get to grips with essential vocabulary for the journey by surveying perspectives on demand and difficulty.

The terms “demand” and “difficulty”, are fundamental to this study. The purpose of this chapter is to explore the ways in which they have been used in academic and professional writing, and to offer clarity about the specific ways in which they will be used in this study.

Baird *et al.* provide the following working definitions:

‘Demand is... essentially the height of the hurdle that the examiners set when they define the syllabus and question papers. Difficulty is how well the students are rewarded for their efforts and tends to be measured quantitatively, for example as an average score’ (Baird, *et al.*, 2009, p. 7).

Within the definition, demand is fixed at the point of question setting, whereas difficulty varies between students.

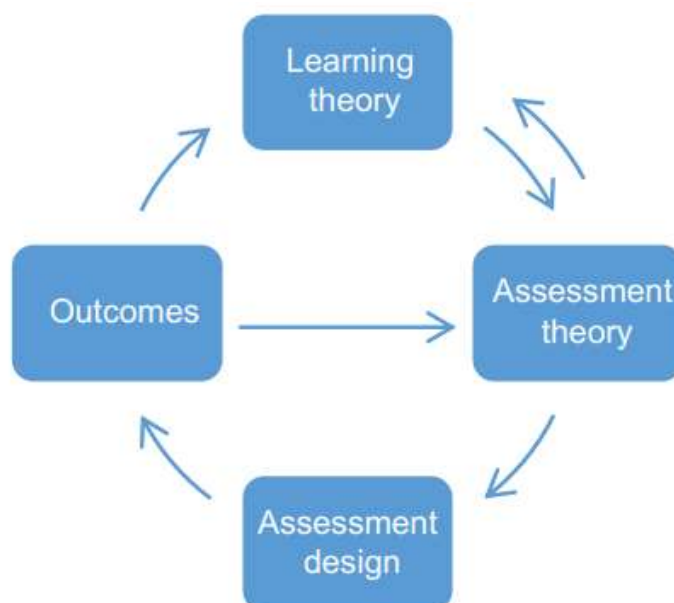
To clarify this distinction, Baird *et al.*, (2009), ask readers to imagine a situation in which two groups of students from two different schools sit the same examination. In one school, students have been taught the topic (Pythagoras’ theorem, say), whereas in the other they have not. When both groups of students encounter an examination question that requires a knowledge of Pythagoras’ theorem, the cognitive demands made by the question are same for both sets of students, but the students who have not studied the topic will, other things being equal, find the question considerably more difficult. As Baird *et al.*, summarised:

‘Difficulty is the quantitatively measured performance of students, whereas demand is the cognitive load of a topic. Student preparation, in the form of tuition or other study, separates demand and difficulty. However, demand and difficulty are often not distinct in students’ experiences’ (2009, p. 7).

Demand ought, by these definitions, to be in the control of examiners. Already there are potential problems with this idea. While it appears reasonable to assume that a question, before it is addressed by a student, has a certain (fixed) level of demand, is it necessarily the case that this means that the 'cognitive load' it imposes is the same for each individual student? As Baird *et al.*, (2009) point out, some demands are not always intended. And some features of a question – the wording, for example – may affect some students more than others. As is subsequently discussed in Section 2.1, some examiners may be unaware of, or may not intend to use, some of the sources of demand that nonetheless feature in their questions. Furthermore, if 'demand = cognitive load' then this may vary between students, depending on the technique the student uses to approach the question, and how the cognitive load is made up. Cognitive load will be discussed in detail in Chapter 3.

In a more recent study, Baird, Andrich, Hopfenbeck, and Stobart (2017) set out an idealised view of the possible relationship between theory and assessment design, (as shown in Figure 2). They argue that a rational observer might anticipate that theories of learning would influence assessment theory and assessment design, and that theories of learning and assessment have reciprocal influences on each other. For some, this appears self-evident; for others, assessment is quite separate from learning. Baird *et al.*'s position is that 'not only does there need to be a correspondence between learning and assessment theories, but that it should be stronger than it has been to date' (2017, p. 318). The present study makes this link directly, focussing on the correspondence between learning theory, (at the top of Figure 2), and assessment design, (at the bottom of Figure 2).

Figure 2 - Idealised Relationship between Theory and Assessment Design



Source: Baird *et al.*, (2017), p. 318.

This model of the idealised relationship between theory and assessment design misses one crucial aspect of the system: the *student*. Yet students are not only the focus (and consumer) of assessment design, they are also its majority stakeholders because the outcomes of educational assessments arguably affect their life chances (Taylor, 2005) far more than they affect teachers, examiners, or researchers. This significant omission appears to be a feature of much assessment-focused research: it is, as Wood (2007) infers, all too easy to lose sight of students in the assessment system. It is this omission that this study seeks to address by seeking, reporting and critically evaluating the views and concerns of students.

2.1 Demand in examination board literature

Lord Dearing, in his review of qualifications for 16-19-year-olds, proposed 'a coherent national framework for Great Britain that covers all the main qualifications and achievements of young people at every level of ability' (Dearing, 1996, p. 1). Central to his work was his understanding of the role that examinations play within the assessment and awarding of those qualifications.

Dearing understood that ‘an examination is only as difficult as the questions and mark schemes from which it is built up’ (Dearing, 1996, p. 1). As Fischer-Hoch and Hughes noted (1996, p. 1), this remark pinpointed ‘what should be the key focus of research in examination difficulty: the question and associated mark scheme.’ A practical realisation can be seen in the need for examinations to comply with guidelines from the School Curriculum and Assessment Authority’s (SCAA²³) Mandatory Code of Practice for GCSEs (1995):

‘The question paper must discriminate effectively among candidates... and GCSE papers at the highest tier must provide a suitably demanding challenge for the highest grade to be awarded’ (SCAA, 1995, p. 2).

Though SCAA was relatively short-lived (1993-1997), the clarity of its guidelines is useful. This is because in its code of practice, the concept of demand or challenge is essential to enabling the creation and evaluation of examination questions and papers that fulfil one of their basic functions: discriminating between candidates of differing levels of expertise. Thus, this concept allows the performance of candidates to be recognised and compared, and also enables examinations to be evaluated and compared. Moreover, it allows subjects and even individual examination papers to be ranked relative to others. It can also be seen to allow for complex questions about assessment to be posed and discussed. In this context, it is possible to ask a number of searching questions in relation to demand and difficulty, including how “hard” one subject is, compared to another, and why one year’s examination paper appears to be more difficult than the paper from another year. He *et al.* (2015) investigated different models of effective assessment for students aged 16 in England; they concluded that tiered papers, as in mathematics, brought several advantages, in terms of demand and difficulty:

²³ The School Curriculum and Assessment Authority (SCAA) was ‘formed as result of the [UK Government’s] Education Act 1993 to take over the responsibilities of both the National Curriculum Council (NCC) and the School Examinations and Assessment Council (SEAC), thus drawing together curriculum and assessment functions in one body. It was merged with the National Council for Vocational Qualifications in 1997 to form the Qualifications and Curriculum Authority, this time signalling a convergence between general or academic education and vocational education and training’ (Wallace, S. (ed.) *A Dictionary of Education*, Oxford, 2009, online) <https://www.oxfordreference.com/view/10.1093/oi/authority.20110810105812234> accessed 18.04.2021

'Tiered papers are targeted at the appropriate level of demand and difficulty so that all pupils should have a satisfying experience. This approach can maximise the opportunity for positive achievement for all pupils. Pupils at different ability levels are given a reasonable chance to demonstrate what they know, understand and can do. However, this model relies upon teachers' ability to accurately predict pupils' potential examination performance' (He *et al.*, 2015, p. 84).

There is, however, a basic assumption that underlies the whole formal assessment and examination system: that there is a uni-dimensional property such as "ability in mathematics" which operates across different topic and questions, and which can be measured on a single scale to produce a graded outcome. This assumption is problematic, to say the least. The present study operates within the formalised assessment system of examinations established in UK schools, and its overall purpose is not to recommend the abolition of this system, but rather to suggest ways in which it might be improved. Nonetheless, it is important to discuss and understand this underlying and problematic assumption. Fundamental to the complexities of educational testing, as Koretz states, is that,

'Test scores usually do not provide a direct and complete measure of educational achievement. Rather, they are incomplete measures, proxies for the more comprehensive measures that we would ideally use but that are generally unavailable to us' (Koretz, 2008, p. 9).

Test scores are necessarily incomplete, Koretz explains, because they can test only a small subset of the content knowledge within a curriculum, and because they test only small samples of cognitive behaviour. But we then use these samples to generate estimates (which we call "test results" or "exam grades") of students' mastery of large domains of knowledge and skill. As a society, "we" are accustomed to accept the reporting of a single grade as being an estimate of a student's expertise in a subject (mathematics, say, or English). We have no problem with distinguishing between "expertise in mathematics" and "expertise in English," because they are reported as distinct grades, but we do not expect separate reports within the single grade for mathematics for "mastery of arithmetic," "expertise at logical reasoning," or "facility in remembering and applying a formula," for example. Bringing this back to demand and difficulty, a single examination question may present demands in several different areas of

cognition. The cognitive load of these different elements may not be the same for all students, as discussed in section 3.1.4. Acknowledging these discussions, and without resolving them at this point, the notion that ‘demand = cognitive load’ is nonetheless a useful rule of thumb at this stage in our study.

Looking more broadly, across whole subjects rather than at the individual question level, Coe *et al.* used a number of different statistical methods to investigate ‘whether examinations in some subjects at GCSE could legitimately be described as “harder” than those in other subjects’ (2008, p. 2). Part of their motivation was to discover whether “STEM” subjects – science, technology, engineering and mathematics – were harder than other subjects. They found evidence of high levels of consistency in estimates of subject difficulty across methods and over time. They demonstrated that, at GCSE level, mathematics was generally harder than some other subjects. Specifically, they found that, across an average of five different statistical analyses, mathematics was similar in difficulty to music but more difficult than other arts subjects (fine art and drama), English, technology and physical education, but less difficult than sciences (joint science, biology, physics and chemistry), humanities (geography and history) and modern foreign languages (Spanish, French and German). Coe *et al.* found that mathematics was a little more difficult (relatively) for female than for male students, but that there were negligible differences in terms of students’ backgrounds (disadvantaged/non-disadvantaged) or the nature of their schools (independent/maintained). This interesting quantitative approach has not been replicated since 2008, unfortunately, so its applicability to reformed GCSEs (awarded after 2017) cannot be stated. As Coe and other authors have noted, however, examination performance is affected by many factors apart from question demand, such as motivation, candidates’ intrinsic interest in the subject, the quality of teaching experienced, candidates’ levels of examination preparation, the amount of curriculum time devoted to the subject, and so on (Coe, 2007; Coe *et al.*, 2008; Alton and Pearson, 1995; Goldstein and Cresswell, 1996; Newton, 1997).

The terms demand and difficulty have acquired distinct meanings among researchers and examiners who routinely discuss high-stakes²⁴ (and other) assessments and examinations. They are also often used in ordinary speech and in the wider academic and professional literature in manners that suggest that they are either interchangeable or mean the same thing. Often, the term “difficulty” is used as a catch-all term, covering demand as well as difficulty. Sometimes even examiners engage in this slipshod usage, as shown later in this section. Commenting further upon such issues, Pollitt *et al.*, observed that:

‘Examiners and many varieties of commentator have long talked about how demanding a particular examination is, or seems to be, but there is not a clear understanding of what demand means nor of how it differs from difficulty’ (2007, p. 166).

They went on to give working definitions. Pollitt *et al.*, (2007, p. 196) defined demand as:

‘Separable, but not wholly discrete, skills or skill sets that are presumed to determine the relative difficulty of examination tasks and are intentionally included in examinations/assessments.’

This much is clear: demand intentionally creates difficulty. The demands Pollitt *et al.* mention are among the approaches that examiners can use to increase the cognitive load of a question, for example by increasing the number of steps that a candidate is required to perform, or increasing the amount of information that they are required to hold in their working memory, or increasing the abstraction of the context. These are the tools of the examiners’ trade. Through application of these tools, it is possible to ask more demanding or less demanding questions on the same topic.

²⁴ Assessments are routinely used in schools to measure the learning of students. High-stakes testing is based on the premise that learning will increase if educators (and students) are held to account for the results. By definition, the stakes become high in tests and examinations where results are used to make decisions about students’ progress, admission to college/university, graduation and job prospects; or to determine teachers’ and leaders’ promotion and salaries. For a fuller discussion of high-stakes testing and its effects, see, for example, Jones and Ennes (2018).

Dhillon and Richardson, researchers with AQA²⁵, recognised that ‘the process of understanding the question’ – that is, comprehending and overcoming the question’s demands – ‘consists of numerous mental operations that contribute to the cognitive demands and conceptual complexity of the question’ (Dhillon and Richardson, 2003, p. 2). The terminology they used is, however, a little confusing: they refer to ‘intrinsic question difficulty’ and ‘surface question difficulty’ (2003, p. 4), although in both cases they clearly mean demand (in our definitions), since these are features that are controlled by examiners. Citing Ahmed and Pollitt (1999), Dhillon and Richardson (2003) likened these numerous demand operations to “hurdles” that the student must overcome in order to formulate an answer, whilst the “number” and “height” of the hurdles contribute to the overall demand of the question.

2.2 Difficulty in examination board literature

Difficulty, on the other hand, can only be determined once an examination has been taken. As described by Pollitt *et al.*, difficulty is,

‘A statistical measure that indicates how likely it is for any given student to score marks, estimated by considering the scores of actual students’ (2007, p. 196).

In other words, difficulty for any given candidate is an empirical measure of the probability that the (given) candidate has of correctly answering the question, compared to all other candidates.

For the inferences made from an examination to have a good level of validity, (as subsequently discussed in Section 3.3), the examination needs to be a fair test of the knowledge, understanding, and skills of the candidate, whilst the examination result needs to stand as a good predictor for the candidate’s future success in the subject. This is a heavy burden for any statistical measure to bear (see Koretz, 2008), and it is part of the reason why this study believes that examination questions need be scrutinised and, where possible, improved. Boud

²⁵ AQA, formerly known as the Assessment and Qualifications Alliance. This is one of the three major examination awarding authorities operating in England.

argued for the 'double duty' of sustainable assessment (making an analogy with sustainable development), where assessment 'meets the needs of the present without compromising the ability of the students to meet their own future learning needs' (2000, p. 151). This study takes a small step in this direction by encouraging students to reflect on their own assessments.

Pollitt *et al.*, (2007), were clear in their view that an examiner ought to be able to estimate beforehand the demands that their examinations (and individual questions) place on candidates; and that examiners ought also to be able to determine afterwards the actual difficulty that the examination (and individual questions) posed for those same candidates.

The difference between demands and difficulty therefore involves timing:

'While it may be worth asking judges how difficult they think a paper is, this is not the same as asking how demanding it is; the former is a prediction, the latter a judgement' (Pollitt *et al.*, 2007, p. 196).

In an ideal world, there would be no difference between the examiners' prediction of the demands of their questions and the subsequent judgement of difficulty derived from the performance of candidates. In practice, however, it is clear that this is not usually the case; this is shown by the numerous changes made to grade boundaries each year by examination boards in the UK. It follows that examiners may have an incomplete or imperfect view of the ways in which demand in an examination question creates difficulty for candidates. If there are elements governing question demand that are unpredictable or not within the control of examiners, then this potentially calls into question the validity of the purposes and interpretations to which the examination grades may subsequently be put.

Jackson and Lismore-Burns (2012) examined the difficulty of questions in four different topics within GCSE Mathematics.²⁶ Without defining 'difficulty', the authors observed that:

'It is important to note that there are two potential sources of difficulty in an examination. The first is the difficulty of the topics; the second is the difficulty of the specific questions asked. It would, for example, be possible to ask an easy question

²⁶ This study was published by the research arm of AQA. It makes interesting points but it also raises questions.

about a difficult topic, or a difficult question about an easy topic. For that reason, the topics selected for this report were selected because the format of the questions had been consistent over time' (Jackson and Lismore-Burns, 2012, p. 1).

It may be meaningful to talk of the difficulty of different topics, but when examiners and researchers talk of asking 'a difficult question about an easy topic', they are discussing demand and not difficulty. It is just possible that a question may be poorly worded, as subsequently exemplified and discussed in Chapter 6. The notional examiner may have intended to set an undemanding question, but some quirk of presentation or context has meant that candidates have been confused and, since too few answer correctly, it turns out to be a difficult question. This is not what is being discussed here, however; rather, Jackson and Lismore-Burns were referring to the possibility of setting a demanding question on a straightforward topic, or *vice-versa*. Although it is important to note that their study does not state their understanding explicitly, Jackson and Lismore-Burns were here asserting that the difficulty of a question is expressed as a simple function of the proportion of candidates who answered it correctly.

If this proposition is accepted, then it follows that the expertise of a candidate can also be simply defined: a more expert²⁷ candidate is the one who answers more questions correctly. This reflects the way that public examinations work: candidates who score more marks obtain higher grades, and these grades are generally recognised as reflections of their expertise. Critically, however, this calculation of expertise takes little account of the demand of the individual questions answered. Two candidates might have very different levels of actual expertise but could – at least in theory – gain the same number of marks, if they correctly answer different combinations of more and less demanding questions. In practice, the more demanding questions usually have more marks allocated to them, to offset this problem, and it is rare for an able candidate to answer incorrectly questions that are less difficult, although it could happen if questions were worded in confusing ways.

²⁷ The words 'expert' and 'expertise' are preferred in this study to the more problematic and nebulous terms 'able' and 'ability.'

Jackson and Lismore-Burns used diagrams to explore the relative difficulty of GCSE mathematics examination questions (see, for example, Figure 3). They explained their method simply:

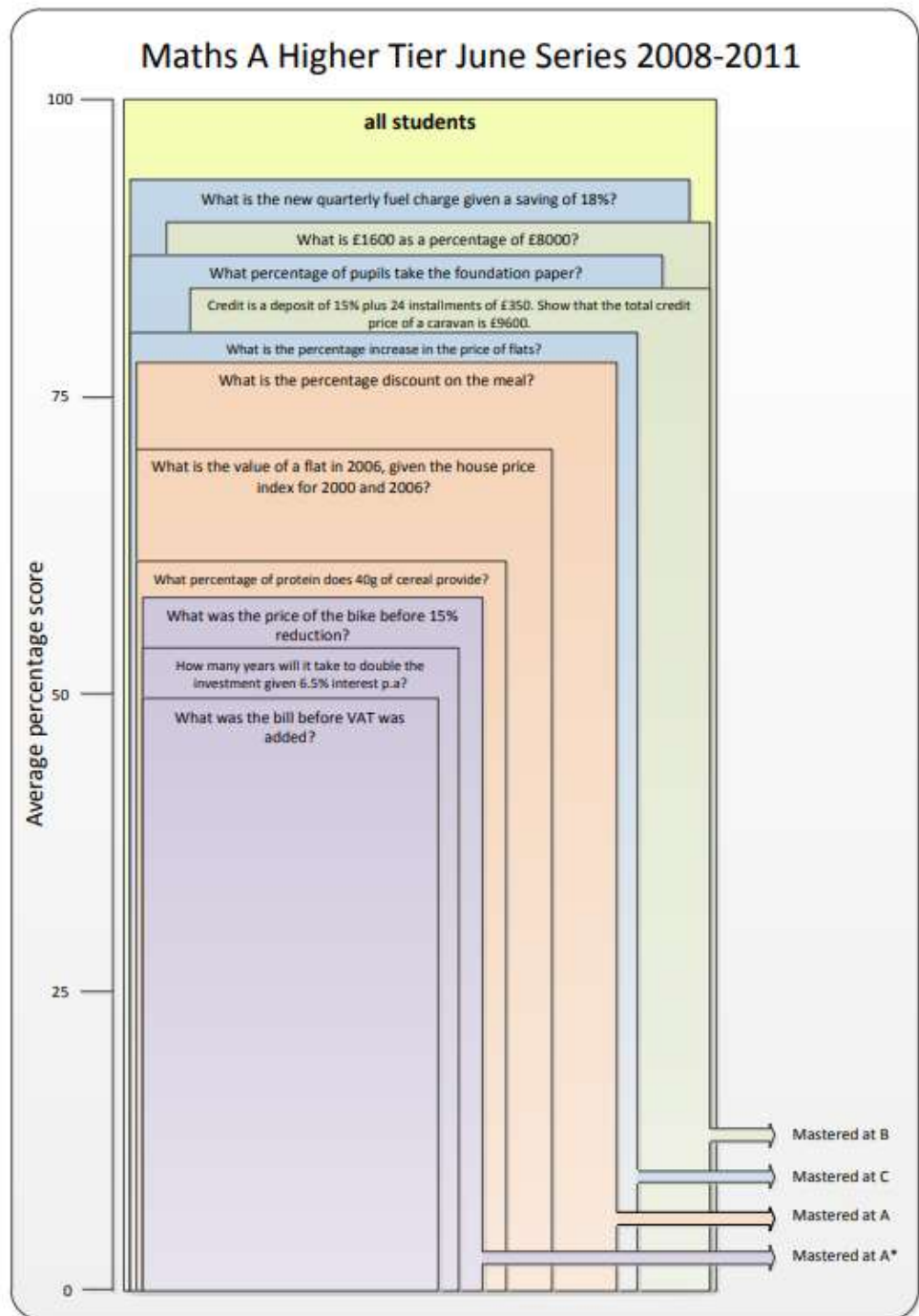
[The figures] 'represent the difficulty of questions within each of the four topics selected, and the level of student ability, represented by grade, at which mastery of each question was achieved. The height of each question box relates to the average percentage score of the candidates in that cohort. Thus, the lower the height of the box, the more difficult the question. The colour of the question box represents the grade at which the question appeared to have been mastered. So, for example, in Figure 2 on quadratic equations, the question 'factorise $x^2 + 3x$ ' was less demanding for candidates, with a mean percentage score of 72%, thus it has a greater height. It was mastered at grade B. The question 'solve $2x^2 + 3x - 7 = 0$ ' was more difficult for candidates, with a mean percentage score of 25%, so the height is much lower. It was mastered at grade A*' (Jackson and Lismore-Burns, 2012, pp. 1-2).

Jackson and Lismore-Burns make reference to the GCSE grading system. When GCSE examinations were first introduced in 1988, grades were awarded with letter names, A to G, where A was the highest grade. This system was similar to grades A to E of the O level examinations that GCSEs replaced, but with additional lower grades of F and G. Grade U (unclassified) was given for work that fell below grade G, and such work was not certificated. In 1994 an additional high grade, A*, was introduced for exceptional achievement. Within mathematics and science, different tiers of entry were created in GCSE. On first introduction, mathematics had three tiers: Foundation (grades G, F, E and D), Intermediate (grades E, D, C and B) and Higher (grades C, B, A, and subsequently A*). The intermediate tier was scrapped in 2017. When GCSEs were reformed, with first awards of the revised system being made in 2017, the grading system was also changed, to numbers 9 to 1, with 9 being the highest.

The first of the diagrams presented by Jackson and Lismore-Burns is shown in Figure 3. The height of the coloured block corresponds to the average proportion of students who answered the question correctly (the width of the coloured block appears to hold no significance: it is just a scaled proportion of the block's height). In this way, Jackson and Lismore-Burns mapped the difficulty of GCSE mathematics examination questions. Unfortunately, they gave no

commentary to accompany their analysis: they presented no discussion of what underlying concepts might be represented by their difficulty charts, and nor did they give any explanation of what they meant, for example, by 'mastered at grade B.' From their analysis, it is not possible to untangle the difficulty of the topic – as experienced by the students – from the demands of the question, and they themselves made no attempt to do so. It is possible to infer that 'what is £1600 as a percentage of £8000?' (green = 'mastered at B') is a less abstract question and requires less use of working memory than 'what was the price of the bike before 15% reduction?' (mauve = 'mastered at A*'), which may suggest that the mauve question imposes higher levels of cognitive load demands than the green question, but Jackson and Lismore-Burns did not make this step in their analysis. This is deficiency which, once more, points to the need for the current study.

Figure 3 - Relative Difficulty of Questions in GCSE Mathematics



Source: Jackson and Lismore-Burns (2012, p. 3).

Nevertheless, Jackson and Lismore Burns (2012) did make a series of recommendations, which suggest that, at least in 2012, the field was somewhat under-studied:

‘Following this initial exploratory study, it would be useful to consider issues such as:

- To what extent do the teachers’ perceptions of the specific difficulties faced by their candidates align with the outcomes from this analysis?
- To what extent are the assumptions of key stakeholders about what students at different grades are likely to know, understand and be able to do, supported by the outcomes from this analysis?

A more comprehensive review, incorporating additional data and seeking to understand in more detail the specific underlying difficulties experienced by candidates would clearly be of interest here’ (Jackson and Lismore-Burns, 2012, p. 2).

It is interesting that Jackson and Lismore-Burns would enquire into the ‘perceptions of difficulty’ of the teachers and other ‘key stakeholders,’ but not ask the students themselves. In another paper, also published by the research arm of the examination board AQA, Spalding (2011) noted that question papers tend to have an incline of difficulty, that is, they begin with less difficult questions and proceed to more difficult questions. Various reasons for this are suggested, including not wanting to increase candidates’ anxiety at the start of an examination. A slightly naïve tone exudes from the papers of Spalding and of Jackson and Lismore-Burns, and suggests that the science and craft of constructing examination questions and papers was rarely considered and little understood in the early 2010s by examiners at one of the world’s largest English-language examination boards. The tone of Spalding’s remarks, on the lack of guidance given by AQA in relation to the positioning of more and less difficult questions, for instance, indicates that this issue was not considered to be of any great professional concern:

‘Currently, AQA does not provide question paper writers with procedural guidance for structuring papers. However, unwritten rules of good practice have evolved over the years and most of AQA’s papers do naturally follow the guidelines outlined by the research literature’ (Spalding, 2011, p. 13).

Spalding observed that ‘the difficulty of optional questions is unpredictable even to experienced examiners’ (2011, p. 13), who were accustomed to using sophisticated statistical

techniques such as Item Response Theory analysis. It can be inferred from this that the science of determining the difficulty of specific questions was little understood.

There also appears, from Spalding's paper, to have been some confusion from an examination board employee around the distinction between demand and difficulty. Spalding used the term difficulty in two different ways. She wrote that 'the questions were carefully selected so as to assess the same skills and to be of equal difficulty' (2011, p. 14), using the term to describe aspects of demand (not difficulty), since these questions had not yet been encountered by real students. It is not clear whether Spalding was noting that the demand of apparently similar questions – a function that, as Pollitt *et al.*, (2007) showed, ought to be predictable and known by examiners – was not in fact accurately predicted. Or perhaps she was saying that even experienced examiners, having made predictions of the demands of similar questions, were confounded by the outcomes that these questions created when tested on real candidates in examination conditions. If 'even experienced examiners' (Spalding, 2011, p. 14) could not (and arguably still cannot) accurately predict the demands of the questions they create – that is, if predicted demand does not reliably translate into experienced difficulty – then the extent to which they can claim that the results of their examinations are robust and reliable indications of students' expertise is questionable. This, in turn, may threaten the validity of inferences made from the results of such examinations.

It may be desirable for examiners to remove unnecessary detail in questions, in order to ensure maximum accessibility, an issue that will be discussed in terms of cognitive load in section 3.1.2. Commenting further upon this, Dhillon and Richardson (2003) noted that the relationship between accessibility and demand is not straightforward. Their study is a rare example of an investigation from exam board researchers of the relationship between demand and difficulty; there was some involvement of students in that a few comments were sought and reported; some of the conclusions were not as expected, which made for better

discussion. However, their use of the terms demand and difficulty was insufficiently distinct, confirming the need for shared understanding of these common terms that this thesis argues.

Dhillon and Richardson found, counter-intuitively, that their attempts to simplify a problem had, in some cases, actually made them harder for students. They plotted performance scores (which are the complement of difficulty) against the results of attempts to manipulate intrinsic demand. For most questions they found that, as expected, as intrinsic demand increased, performance decreased, that is, the more demanding the question was, the less successful the students were. This pattern was disrupted, however, for some items that had been highly scaffolded²⁸. Increased scaffolding was intended to reduce the cognitive load, which should have resulted in the performance score being increased. Instead, they found that the high levels of scaffolding created questions that were harder and that, accordingly, performance scores decreased. Conversely, questions with low scaffolding were the ones that students answered most successfully.

Dhillon and Richardson commented that, for 'with respect to scaffolding effects the remaining picture runs entirely counter to predictions' (2003, p. 5). Rather than helping the students, Dhillon and Richardson suggest that scaffolding in fact hindered them 'by providing too much information, too densely presented, or... information they found distracting' (2003, p. 6). This suggestion was corroborated via student comments. One noted that the additional details 'were a bit distracting, it pulls you away from the question,' whilst another said that 'you lose your train of thought because you have to think of several things at once' (2003, p. 6). It follows that steps intended to increase or reduce demand do not always have the intended

²⁸ The metaphor of "scaffolding" was introduced by Wood *et al.* (1976) to explain how teachers offer support to make learning tasks more manageable for students. This approach has been challenged (by, among others, Smagorinsky, 2018) as being too general and unfocused, excusing teacher interventions that are not in the learning interests of the student. However, as Anghileri explains (2006, p. 33), 'despite problems, this metaphor has enduring attraction in the way it emphasises the intent to support a sound foundation with increasing independence for the learner as understanding becomes more secure.'

consequences, and also that students are able to provide evidenced feedback that could improve examiners' understanding.

In a paper for the Centre for Education Research and Policy (CERP), Dhillon (2003) further explored the concepts of demand and difficulty from an examiner's perspective. Citing Pollitt *et al.* (1985) and other authors, she distinguished between "concept", "process", and "question" difficulty. Concept difficulty is 'concerned with the inherent conceptual complexity of the subject matter and is determined by the degree to which the concepts involved in a question are abstract or concrete', whereas process difficulty 'concerns the difficulty of the cognitive operations and the degree to which they utilise finite cognitive resources' (Dhillon, 2003, p. 2). Dhillon preferred the more inclusive term of "intrinsic difficulty" to encompass and replace these earlier definitions. Similarly, Dhillon coined the term "surface difficulty", collecting together the format-bound 'linguistic and structural properties of the question and the appropriate use of mark schemes' (2003, p 2). (According to the definitions already advanced by Baird *et al.* (2009) and Pollitt *et al.* (2007), these matters should all be labelled "demand", not difficulty, since they are in the purview of examiners setting questions.) Pollitt's distinction between "legitimate" (intended and justified) and "illegitimate" (unintended) demand is at the heart of examiners' concerns. Fischer-Hoch and Hughes (1996), labelled these as "valid difficulty" and "invalid difficulty" respectively. Dhillon summarised this discussion:

'Legitimate sources of question difficulty are those that intentionally and transparently seek to assess skills or knowledge representative of a level of aptitude or proficiency in a subject. Conversely, illegitimate question difficulty is indicative of a communication failure between two or more of the 'characters' in the assessment dynamic; the question setter, the candidate and the marker' (Dhillon, 2003, p. 3).

This is a significant comment, as it brings in the candidate as the third essential side of the assessment triangle. Too often, as has been seen, the voice of the student or candidate is silent or missing, and their role and function is overlooked. A missing element, however, is any discussion of the threat to validity of these unintended aspects of demand.

Commenting further on the impact of unexpected demand factors, Baird and Black observed, in the context of a discussion around test theories, educational priorities and reliability of public examinations, that, because public examinations are grounded in a particular syllabus, their main aim is,

‘To find out about whether students know this stuff. Therefore, there are no a priori expectations necessarily about how the total scores or item difficulties will be distributed... In fact, predicting the difficulty of items turns out to be a very difficult job, even for experts (Cresswell, 1997, 2000; Impara and Plake, 1998), so there are surprises in examination results about which questions were most difficult and the shape of the total score distribution’ (Baird and Black, 2013, p. 11).

Baird and Black here contextualised the demanding task that examiners face in pre-determining the cognitive load of both individual questions and entire examination papers.

It is appropriate for examiners to remove barriers to students’ understanding questions, but it would not be appropriate for them to adopt teachers’ scaffolding approaches within the structured testing of an examination. Crisp and Macinska state that:

‘The aim of improving the accessibility of a question is not to reduce its demands but to provide students with a better opportunity to demonstrate their knowledge and skills by removing any obstacles to question comprehension’ (2020, p. 2).

As Oates noted, ‘improving accessibility’ should not be a ‘pre-eminent concern in assessment’, since

‘Pursuing some accessibility aims can have a very specific, adverse impact on standards of demand...The best policy scenario is that the tension between enhanced accessibility and maintenance of standards is held in careful balance’ (Oates, 2020, p. 1).

In this way, Oates echoed the experimental findings of Dhillon and Richardson (2003) and others, who showed that attempts at manipulating accessibility can have unpredictable, unintended and possibly undesirable effects. His conclusion, therefore, was that, although the goal of improving accessibility is certainly desirable within examinations, this must not be at the expense of making questions suitably demanding for students. Scaffolding is for teachers, and it aims to lower cognitive demands for the purposes of learning. Improving accessibility is

for examiners, and it must not lower the demands made on students in examinations, or else the examination will discriminate poorly between students of different levels of expertise. Ahmed and Pollitt (2007) investigated the effect of making questions more focused or less focused. They found that, in the context of the performance of test items, more focused questions proved better than less focused ones: they improved accessibility without reducing demand.

Crisp and Grayson applied a cognitive model, based on Pollitt and Ahmed's (1999) generic psychological model of the question answering process, to identify the features of the A level physics examination questions 'that could have potentially influenced their difficulty' (Crisp and Grayson, 2013, p. 352). They deployed regression analysis and Rasch statistical analysis techniques, and noted several factors that make item difficulty modelling problematic in UK examinations, including the lack of 'item banking and substantial reuse of items'²⁹ (Crisp and Grayson, 2013, p. 346), which make models of difficulty potentially unstable. Their extensive data set and complex mathematical analysis advanced a range of conclusions. They commented, for instance, that:

'Within the regression model, four question features were found to be significant predictors of question difficulty: total amount of reading, use [of] physics concepts (recall or understand), work with symbols, and intermediate or complex calculations' (Crisp and Grayson, 2013, p. 367).

In other words, examination questions were found to be more difficult if students were required to read more; if they needed to recall or understand subject-specific concepts; if abstract symbols were involved (as in algebra, for example); or if questions required significant calculations to be performed. These conclusions stand as a useful evidence-based confirmation of many teachers' and students' intuitive understanding of question demands and difficulty and relate closely to the thinking around demand of Hughes, Pollitt and Ahmed (1998).

²⁹ "Item banking" is the creation of a bank of questions, the properties and performance of which are known, that can be re-used in subsequent assessments. This works well enough for assessments where the questions are not put into the public domain, so that they are encountered by candidates for the first time in the actual assessment (see Hambleton and Swaminathan, 1985).

Accordingly, it can be concluded that it is not simply the topic demand of questions that increases their difficulty, but that other factors, such as abstraction and length of question, tend to increase the total cognitive load for students. Examiners may wish to increase accessibility without decreasing demand but, given the complexity of overlapping constructs affecting demand and difficulty, there is little evidence to suggest that examiners are certain about how to do this. In any case, few researchers have tested their theories with students.

Crisp and Macinska (2020) demonstrated that the accessibility principles articulated by the examination board OCR (Oxford Cambridge and RSA) for examination questions were appropriate and that they worked in practice with students. Crisp and Macinska sought the views of 57 students who had sat either the original version of questions from OCR's Foundation Tier Science GCSE papers from June 2018 or the final version, once the examination board's accessibility principles had been applied. These students were from a wide mixture of school backgrounds, including comprehensive, independent, and special provision. The researchers concluded that,

'The students' views gathered... suggest that the accessibility principles that we investigated are appropriate and should continue to be applied to help ensure students can understand and access future exam questions' (Crisp and Macinska, 2020, p. 9).

This is a rare example of an examination board seeking validation of their methods from students who had sat their papers. Regrettably, the methodology of the study was not as rigorous as that of many others from Cambridge Assessment's researchers: no empirical data was gathered, no statistical analysis was applied, and although students' views were reported – mostly as percentages preferring one version of the question or the other – there were no verbatim comments. This lack of rigour may affect the validity and applicability of their findings; no similar studies have been carried out to establish if these findings are reliable. Although students were involved in the study, opportunities to strengthen the validity of the findings, for example through the inclusion of their authentic voices in quotations in the report, were not taken.

As shown in this section, some education professionals connected with examination boards have engaged, in theory and through practice, with the connection between demand and difficulty in examination questions. They found that it was possible to reduce the cognitive load demands by adjusting aspects of the question, but that the effects of these manipulations were not always predictable. In the majority of cases they did not, however, follow through and investigate the impact of changes in question design on the students who answer the questions. Researchers who did involve students in their studies tended to be those who were interested in improving the accessibility of examination questions.

2.3 Demand and Difficulty: Conclusion

From the works surveyed in this chapter, it is possible to summarise the definitions of demand and difficulty. Demand is the cognitive load of a topic; it is determined by examiners; and it is measured qualitatively. Demand ought to be in the control of examiners, but some demands are not always intended. Difficulty is the quantitatively measured performance of students. The same demands in a question may result in different levels of difficulty for individual students, depending on their individual levels of expertise and preparedness. These are the ways that the terms demand and difficulty will be used in this study.

It may be further concluded, with reference to examiners' and examination boards' professional involvement in the concepts of demand and difficulty, that there has been an increasing interest from research professionals allied to examination boards, since the beginning of the 21st century, in investigating the relationship between question demand and the difficulty that students experience. Researchers have found that it is possible to manipulate the demands of questions in ways that directly affect the difficulty experienced by students, and they have been frequently able to anticipate correctly the ways in which demand has affected difficulty. Questions were found to be more difficult if students were required to read more, if they needed to use more specialist knowledge recall, and if they

presented higher levels of abstraction. In a small number of cases, however, increasing the help given to students by adding additional detail intended to reduce the demand of questions actually had the opposite effect, and made the questions more difficult. Researchers had not predicted these effects. The unpredictable effects of some question manipulations undermine the ability of examiners to set questions of reliably predictable demand. This constitutes a threat to the validity of inferences made from examination results. The lack of item banking or systematic re-use of examination questions make the modelling of question difficulty unstable and unpredictable in the UK.

Students were involved in only a small number of research studies into demand and difficulty that were carried out by researchers associated with examination boards. In the few studies where students' views were sought, they were not reported verbatim or in detail, which meant that the contribution students might have made to the development of examiners' understanding was diminished considerably. It follows that, without the involvement of student feedback in evaluating how well examination questions work, or revealing the hidden pitfalls of various features of question design, the link between question demand and difficulty will remain not only unpredictable but also a threat to validity. This present study therefore creates and implements suitable research instruments, producing and interpreting data that is to some extent quantitative but mostly qualitative, to bring student voices into research literature concerned with the evaluation of examination questions.

The emphasis in this chapter has been on literature and professional writing that is concerned with demand and difficulty, so as to clarify these important terms. In the next chapter, a wider review of literature is undertaken, reflecting the broad compass of this field of study.

Chapter 3: Literature Review – Mapping the known world

In this chapter I review the “maps” and “travel literature” for demand and difficulty in examinations: cognitive load theory, taxonomies of learning, validity, and student voice. I find that much of the literature is written for the “commercial traveller” and needs suitable translation and adaptation for my purposes.

To a student in an examination, the “view from the desk” may appear restricted, limited by the pressures of contemplating the next 60 or 90 minutes in a sometimes-overwhelming subject-based fug of cognitive activity.³⁰ A student who looks at an examination question in the context of a lesson may take a wider and more relaxed perspective. To the researcher, contemplating the considerable span of literature that bears upon examination questions, the view needs to be more detached still, both calmer and broader (Snyder, 2019). For this study, literature reviewed encompasses four distinct aspects: cognitive load theory; taxonomies of the learning process and student responses; thinking about validity; and a consideration of ‘student voice’ in relation to assessment and research. These four areas are pertinent for review because they all connect with different aspects of the student experience of demand and difficulty in examinations. Cognitive load theory attempts to understand – and taxonomies of learning seek to explain – how human beings develop understanding of subjects taught in schools, and therefore how students can use the same cognitive processes in answering questions. Validity is a central concept in assessment, and a clear understanding is essential to the purposes of this study, as a link is sought between examination outcomes and the inferences drawn from them. Finally, since this study attempts to listen to student voices, a survey of relevant literature will inform the approach to be taken and identify gaps in the existing research. By bringing these four distinct strands of knowledge together, this study evaluates how thinking and understanding in each area contributes to an overall

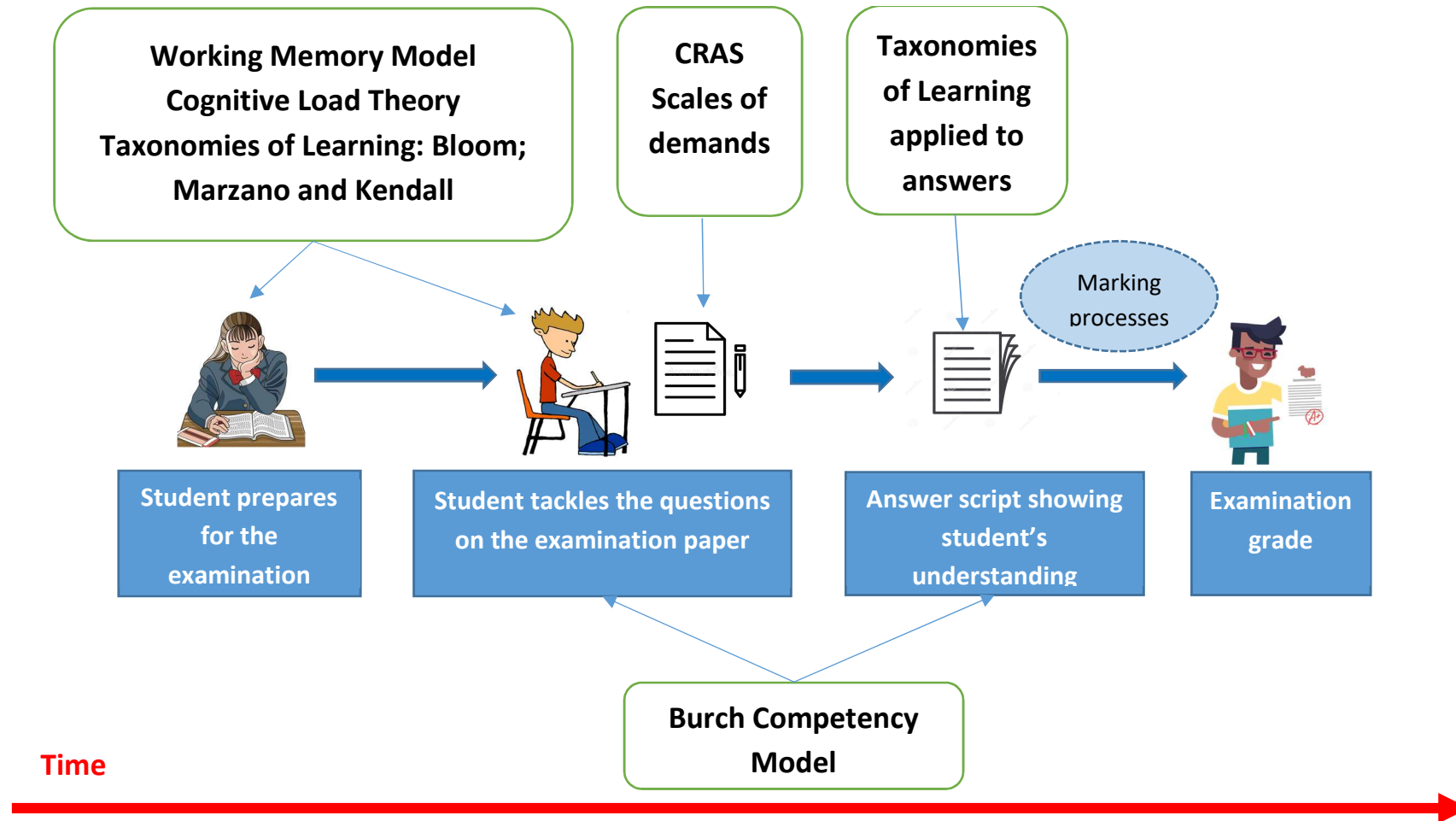
³⁰ The experience of two students preparing for public examinations is given verbatim by Ofqual in this 2019 blog: <https://ofqual.blog.gov.uk/2019/03/08/what-is-it-like-to-experience-exam-stress-a-student-perspective/> accessed 02.05.2022

understanding of demand and difficulty in GCSE examination questions. This study is innovative in combining this range of literature for this purpose.

Because the focus of this study is on the experience and understanding of students, a wide range of different types of literature is included in the review. Research books and journal articles form the core of the reading, and these are complemented, supplemented and challenged by professional literature from sources such as books, professional journals, websites, news media and blogs. Educational thinking is a dynamic field, and many authors – including academics, researchers and teachers – find the immediacy and accessibility of internet posts and printed media beneficial to disseminate, discuss and challenge ideas.

To give the “big picture” for this literature review, the ways in which the different theoretical models work together is shown in Figure 4 below. In this diagram, the interactions and contributions of the different theoretical models used in this study can be seen. Each has a part to play in illuminating the workings of the different parts of the process, as a student prepares for and answers examination questions. Cognitive Load Theory and the taxonomies of learning – Bloom’s Taxonomy, and Marzano’s and Kendall’s New Taxonomy – are models that help the reader to understand the cognitive processes undertaken by a student as they prepare for and sit the examination. CRAS Scales can be used to describe and categorise the different demands imposed by the examination questions. The student’s answers stand as a proxy for their learning, and as such may again be classified and described by taxonomies of learning such as Bloom’s Taxonomy and Marzano’s and Kendall’s New Taxonomy. The Burch Competency Model can provide insight into the student’s apparent level of expertise as they prepare for the examination and, from their answers, as they sit the examination.

Figure 4 - How and where the Different Theoretical Frameworks and Conceptual Models operate in the Present Study



Source: author's own

3.1 Working memory and cognitive load theory, and their relevance to public examinations

John Sweller (2017³¹) notes that, 'without an understanding of human cognitive architecture, instruction is blind'. Indeed, Cognitive Load Theory is an important conceptual framework for understanding educational assessment, as it seeks to rationalise the demands that learning tasks and test items impose on students. For proponents of cognitive load theory, such as Sweller, Chandler, Van Merriënboer and Paas, the fundamental claim is that, if teachers do not understand how thinking and learning works on a basic cognitive level, they will not be able to design teaching strategies that enable learning to take place effectively and efficiently.

In the context of examination questions, students experience a cognitive load when they attempt to answer a question or complete a task (Chandler and Sweller, 1991). Cognitive load theory provides both a conceptual framework and the vocabulary to investigate and describe the demand imposed by learning tasks and examination questions. Cognitive load theory depends on an understanding of the load placed on working memory and the role of long-term memory. This consideration of cognitive load literature therefore begins by exploring how memory works.

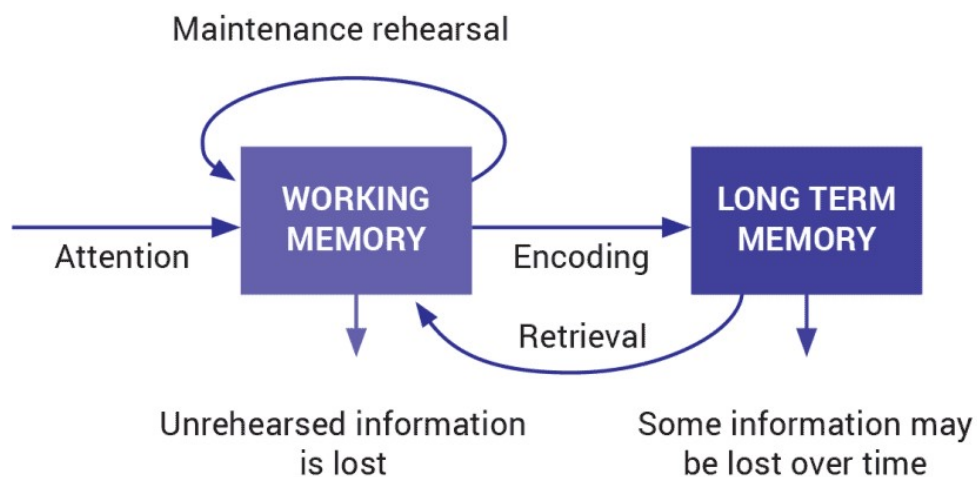
3.1.1 Working memory

As students tackle the 'demands' of examination questions, their working memories are immediately and actively engaged: a clear understanding of the operation of and constraints on students' working memory is therefore important in this study. Case studies of individuals in the 1960s-70s who experienced either long- or short-term memory impairment led to the observation that patients with defective short-term memories were nonetheless able to perform a range of cognitive functions. Atkinson and Shiffrin (1968; 1971) explained that the previously held idea that short term and working memory were one entity no longer seemed

³¹ Sweller was speaking at the ACE Conference/researched Melbourne, Australia, 01.08.2017: <https://www.youtube.com/watch?v=gOLPfi9Ls-w> accessed 02.06.2022

satisfactory. Instead, Baddeley and Hitch proposed a tripartite structure for working memory (1974). In their model, which has been hugely influential in psychological and educational circles, working memory has a coordinating function in cognition, filtering incoming sensory information, storing it temporarily, processing it and enabling the formation of schemas³² that may go on to be stored in the long-term memory (Figure 4). Working memory therefore ‘stands at the crossroads between memory, attention and perception’ (Baddeley, 1992, p. 559). Working memory is the conscious part of human memory; it can handle only a very small number of interacting cognitive elements (Baddeley and Hitch, 1974). ‘Alone, working memory would only permit relatively trivial human cognitive activities. Long-term memory provides humans with the ability to vastly expand this processing ability’ (Paas, Renkl and Sweller, 2003, p. 2).

Figure 5 - Diagram of Human Memory



Source: Likourezes (2021).

The Baddeley and Hitch model of memory shown in Figure 5 can be applied to students learning in school, working from left to right in the diagram. A variety of sensory information – brought to the student’s attention through seeing, hearing and other senses – comes from the outside world into the student’s working memory. If information is not used, it is retained for

³² Chambers Dictionary (1993, 1539) gives a definition of ‘schema’ as ‘a mental picture of a thing in the imagination, which the mind uses to help perceive or understand it more clearly.’

only a short time – between 15 and 30 seconds, according to Atkinson and Shiffrin (1971) – and is then likely to be lost. As information is used and rehearsed, it is encoded into schemas that can be stored in a student’s long-term memory. Sweller, in collaboration with Chen, Paas and Castro-Alonso (Chen *et al.*, 2018), found that working memory is fixed in capacity but, if students are engaged in heavy cognitive effort, working memory is depleted. In addition, although students may expend greater effort in the context of high-stakes examinations, the demands of the examination may defeat them in the face of the twin challenges of high cognitive load and examination stress: Putwain and Symes suggested that, ‘it is possible that the cognitive load on working memory arising from the combination of worry and examination demands may be too high to be compensated by effort’ (2018, p. 482). For students in examinations, then, material from the question enters the student’s working memory from their reading of the examination paper. Understanding, gained from prior learning, is retrieved from the student’s long term memory and combined with the demands of the question in a problem-solving activity. This problem-solving process may undergo several iterations before the student commits to writing an answer on the examination paper.

Baddeley (2002) explained that many traditional teaching methods overload a learner’s working memory, and that this creates a “bottleneck” of the cognitive system. When presented with a new problem, a student quickly scans the problem to see if it matches any of the schema stored in their long-term memory. If it does, then this schema information can be retrieved and used, in the student’s working memory, perhaps combined with the new information presented in the problem, and processed to form a solution.

So far, information and knowledge have been discussed as if they were unidimensional. This is self-evidently an over-simplification. The evolutionary psychologist David Geary drew a distinction between biologically primary and biologically secondary knowledge (Geary, 1995; Geary and Berch, 2016). Biologically primary knowledge is knowledge that human beings have evolved to acquire without needing to be taught explicitly (Sweller, 2017). Examples include:

listening, speaking, problem solving, facial recognition, and generic cognitive skills. Biologically secondary knowledge includes different cognitive skills, such as reading and writing, and domain-specific knowledge, such as mathematics and science – in fact, almost everything that is taught in schools. It might be said, as Sweller has done (2017), that schools have been developed purely to transmit and teach biologically secondary information. It is therefore vital that teachers understand how biologically secondary information is learned effectively: cognitive load theory provides this explanation.

3.1.2 Cognitive load theory

John Sweller has asserted that biologically secondary, domain-specific knowledge is all learned in essentially the same way (Sweller, 2015 and 2017). Although this insight was not part of Sweller's initial work, cognitive load theory now is firmly grounded in evolutionary psychology (Sweller *et al.*, 2019). Sweller's work on cognitive load theory grew from his desire to understand how students acquire domain-specific expertise, as encountered in his investigations into how worked examples helped students learn algebra (Sweller and Cooper, 1985; Cooper and Sweller, 1987; Sweller, 1988). In a highly influential article, Miller (1956)³³ proposed that humans were able to hold only limited amounts of information in their "short-term" memory (as it was referred to at the time), insufficient to explain the management of complex cognitive tasks; De Groot (1966) had observed that what distinguished expert grand master chess players from amateur players was their (long-term) memory of board configurations. Similar findings to De Groot's were obtained in a variety of other domains during the late 1970s and early 1980s: see, for example, Barfield (1986) in information technology; Jeffries *et al.* (1981) in software design; and Sweller and Cooper (1985) in mathematics. These findings led Sweller *et al.*, to the conclusion that,

³³ Miller (1956) has been cited over 37,000 times, according to Google Scholar, accessed 12.03.2022

‘the major factor distinguishing novice from expert problem solvers was not knowledge of sophisticated, general problem-solving strategies but, rather, knowledge of an enormous number of problem states and their associated moves’ (1998, p. 254).

It is not enough for expert chess players simply to remember different board configurations, however: their success lies in their ability to select and recall appropriate prior learning in the form of previously-learned solutions, and apply them to their present situation. Sweller’s work on cognitive load theory, then, explores the interaction between long-term memory and working memory. This is important in closed-book examinations, where students are entirely dependent on knowledge committed to their long-term memory and their facility to process information and instructions in their working memory.

Sweller’s and Cooper’s (1985 and 1987) work with students learning algebra showed them that working memory was incapable of handling highly complex interactions using entirely novel elements, and that inquiry-based or weak, unguided trial-and-error methods were both highly inefficient and usually ineffective for students because they imposed considerable cognitive load. Sweller and Cooper (1985; Cooper and Sweller, 1987) demonstrated that students were better able to address problems that they had not previously encountered if they had been prepared for them through an instructional design that included a series of worked examples. This explicit instruction or worked example approach, where the teacher at first models the approach and then gradually releases responsibility to students, is now commonplace in teaching (see Pearson and Gallagher, 1983, for an early example in the context of teaching reading). It is demonstrated, for example, in techniques such as *‘I do, we do, you do’* championed by, among others, the influential teacher educator Doug Lemov (2010): the teacher models a problem-solving approach; the class and teacher work together on examples of increasing complexity or variety; gradually, agency is delegated to the students, who ultimately solve problems without teacher assistance. Examination questions are commonly used in this way in the classroom: at first, the teacher models the approach, with the intention that, by the time the student sits the examination, they will be able to tackle the question

unaided. The student's memory of the modelled and rehearsed approach is therefore important to their subsequent success, as some students recalled in their responses to the pilot study questionnaire (Chapter 7).

In 1998, Sweller, Van Merriënboer, and Paas co-authored the influential study 'Cognitive architecture and instructional design'.³⁴ In this article, they brought together and expanded upon the three main components of cognitive load: intrinsic load, germane load, and extraneous load. Since, they said, working memory capacity is limited, whereas long-term memory is effectively limitless, holding schemas in different degrees of automation, the assumption must be that, in order for learning to take place efficiently, the load on working memory should be reduced, as far as possible, and schema construction encouraged. Sweller, Van Merriënboer and Paas summarised cognitive load theory thus:

'Its basic premise is that human cognitive processing is heavily constrained by our limited working memory which can only process a limited number of information elements at a time. Cognitive load is increased when unnecessary demands are imposed on the cognitive system. If cognitive load becomes too high, it hampers learning and transfer. Such demands include inadequate instructional methods to educate students about a subject as well as unnecessary distractions of the environment. Cognitive load may also be increased by processes that are germane to learning, such as instructional methods that emphasise subject information that is intrinsically complex. In order to promote learning and transfer, cognitive load is best managed in such a way that cognitive processing irrelevant to learning is minimised and cognitive processing germane to learning is optimised, always within the limits of available cognitive capacity' (2019, p. 262).

Sweller *et al's.*, model of cognitive load, then, involved three elements: intrinsic, germane, and extraneous load. This is how Jessica Mason Blakey, Head of Assessment for Evidence Based Education, defined and explained these three elements, in a blog for Schools Week:

'Intrinsic load is the effort associated with a topic. It is difficult to entirely eliminate this type of cognitive load as a more complex topic will need relatively more mental effort; for example, it's much easier to add 2+2 than solve a quadratic equation...'

³⁴ Sweller, Van Merriënboer and Paas (1998) has had more than 6,900 citations, according to Google Scholar, accessed 25.02.2022

‘Germane cognitive load is the work put into transferring learning to the long-term memory...’

‘The only one that teachers have influence over in the classroom is extraneous load as this boils down to how accessible information is, based on how it’s presented’ (Blakey, 2019).

This seems an over-simplistic explanation, and perhaps illustrates some dangers in boiling down complex research for a generalist educational audience. In section 3.1.4 we will see that the idea that the only influence teachers have is over extrinsic cognitive load is also questionable.

Intrinsic cognitive load may be educationally desirable, as Sweller *et al.*, (1998) suggested, but it also needs to be manageable. To be able to take on a complex task, such as answering a GCSE examination question, a student needs to be able to process several inter-related cognitive elements simultaneously. These elements may be learned separately, but the question cannot be answered ‘until all of the elements and their interactions are processed simultaneously. As a consequence, high-element interactivity is difficult to understand’ (Paas *et al.*, 2003) and it imposes a high cognitive load. For examinations, the focus of this study, intrinsic cognitive load is the element interactivity in the thinking process used to answer the question:

‘Element interactivity is the driver of our first category of cognitive load. That category is called *intrinsic cognitive load* because demands on working memory capacity imposed by element interactivity are intrinsic to the material being learned’ (Paas *et al.*, 2003, p. 1).

In addition to the way in which the desired cognitive elements interact, the ways in which information are presented, and the learning or problem-solving activities required of learners can also impose a cognitive load; ‘when that load is unnecessary and so interferes... it is referred to as an extraneous cognitive load’ (Paas *et al.*, 2003, p. 2).

A third form of cognitive load, germane or effective cognitive load, was also proposed by Sweller to describe the way in which the cognitive load might be initially increased by the

introduction of an instructional (scaffolding) method which, once learned, had the effect of reducing the intrinsic load of the task. Germane load is thus 'the load caused by effortful learning resulting in schema construction and schema automation' (Schnotz and Kürschner, 2007, p. 476). In contrast, Kalyuga argued that cognitive learning theory does not need the concept of germane load, since 'germane load is essentially indistinguishable from intrinsic load' (2011, p. 1). Having shown that the need for germane load came not from a conceptual but from an empirical argument, Kalyuga (2011) analysed a number of empirical studies (DeLeeuw and Mayer, 2008; Schwonke *et al.*, 2011; Gerjets *et al.*, 2004, 2006; Scheiter *et al.*, 2009; Cierniak *et al.*, 2009, all cited in Kalyuga, 2011) and concluded from these that the evidence for germane load was inconclusive at best and confusing at worst. Arguing specifically from a validity perspective, he stated that 'examples of applying similar types of scales for measuring different types of load do not make a convincing case for valid and reliable differential measures of cognitive load' (Kalyuga, 2011, p. 5). Sweller *et al.*, also subsequently reflected that 'germane cognitive load has a redistributive function from extraneous to intrinsic aspects of the task rather than imposing a load in its own right' (2019, p. 264). In this study, therefore, cognitive load is referred to as the sum of intrinsic load and extraneous load, leaving germane load to one side, but with a note that the concept of germane load persists and is attractive to some educators perhaps because it gives a particular value to the role of instruction.

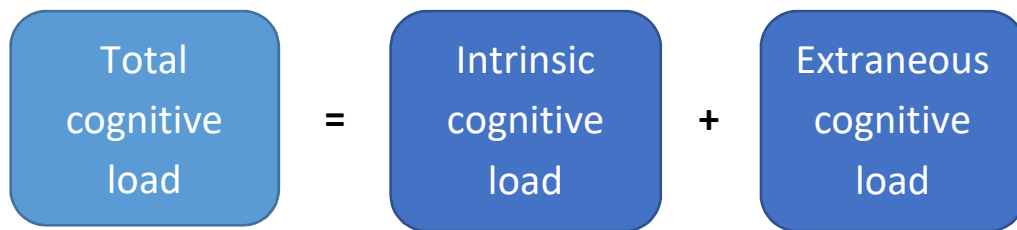
Paas, *et al.* (2003, p. 2) asserted that the different elements of cognitive load 'are additive in that, together, the total load cannot exceed the working memory resources available' if learning is to occur. An increase in a student's expertise, achieved through learning that transfers cognitive elements from the student's working memory to the long-term memory through schema acquisition and automation, reduces intrinsic cognitive load and frees working memory. In this way, a more expert student perceives a task to be less difficult. Between 1998 and 2019, Sweller and his associates explained the workings of seven different effects that could be used to reduce extraneous load and optimise intrinsic load. Four of these effects (the

goal-free effect; the worked example effect; the split-attention effect; and the redundancy effect) are considered within this work when the details of examination questions are discussed. The other three effects (the completion problem effect; the modality effect; and the variability effect) are not addressed as their concerns lie outside the scope of the present study.

Sweller *et al.* also explained that 'the working memory load imposed depends on the number of elements that must be processed simultaneously in working memory' (1998, p. 259). Intrinsic cognitive load tends to be low when only one element at a time is being learned or applied; tasks that require several different cognitive elements to be processed simultaneously impose higher intrinsic cognitive loads. 'Mathematical tasks tend to be high in element interactivity' (Sweller *et al.*, 1998, p. 260) and, therefore, the intrinsic cognitive load that they impose is also typically high. This is an important point for the present study of GCSE examination questions in mathematics because these learning tasks can be very expensive in terms of working memory capacity. It is therefore important to reduce extraneous cognitive load, as far as possible, in order for the task to be accessible to as many students as possible.

Representing this understanding diagrammatically, Figure 6 illustrates how total cognitive load is made up of intrinsic plus extraneous cognitive load. Although simple at this stage, this diagram will acquire additional levels of complexity later in this study.

Figure 6 - Anatomy of Cognitive Load



Source: Author's own (based upon information from Sweller *et al.*, 1998, and Paas *et al.*, 2003).

The discrete contributions of the different elements of cognitive load may initially be hard for teachers, examiners, and students to determine. Sweller stated that,

'While there is a clear distinction between intrinsic and extrinsic cognitive load, from the point of view of a student required to assimilate some new material, the distinction is irrelevant. Learning new material will be difficult if cognitive load is high, irrespective of its source. In contrast, from the point of view of an instructor, the distinction between intrinsic cognitive load is important. Intrinsic cognitive load is fixed and cannot be reduced. On the other hand, extraneous cognitive load caused by inappropriate instructional design can be reduced' (1994, p. 308).

Applying Sweller's insight to examination questions, it may be observed that the intrinsic cognitive load is the thinking load imposed by the deliberate demands of the problem, as intended by the examiners; and that the extraneous cognitive load comprises all the other aspects of difficulty experienced by the student, whether deliberately planned by the examiners or not. The distinction between intended and unintended sources of demand and difficulty is irrelevant to the student: all are experienced as part of the total difficulty of the problem. Yet, to the examiner, this distinction should be vital: in setting questions, examiners should seek to optimise intrinsic cognitive load (that is, the demands) of the question, and to minimise extraneous cognitive load (that is, any unexpected sources of difficulty that may afflict the student). Evidence of this understanding from examiners and examination boards is addressed in Section 2.1. There is an essential link between the effectiveness of students' learning and their ability to reproduce this understanding in the situation of an examination: some students may have been taught in ways that have enabled them to learn more effectively, and some students may have practised more regularly or deployed more effective

revision strategies, that better enable them to recall and use their understanding, so that in the examination they face fewer unexpected challenges to their understanding.

Although cognitive load theory took some time to become established (Mavilidi and Zhong, 2019), it has subsequently become influential among educators seeking to improve the efficiency and quality of teaching and learning. The theory has also been applied to Item Response Theory and other test theories. Schnotz and Kürschner (2007), offered some significant insights that enable more detailed investigation of cognitive loads imposed by examination questions. Like Sweller *et al.*, their starting point was working and long-term memory:

‘Understanding occurs when all relevant elements of information are processed simultaneously in the working memory’ (Marcus *et al.*, 1996, p. 60).

‘Learning is an increase in expertise due to an alteration in long-term memory’³⁵ (Schnotz and Kürschner, 2007, p. 477).

Examiners seek to measure students’ learning: that is, their expertise. In order to locate understanding in the experience of a student, it is necessary to look at some examination questions before considering Schnotz’s and Kürschner’s arguments. There are two contrasting approaches to the construction of questions, such as those that are posed in GCSE Mathematics examinations. The first is to provide a problem, which may be more or less abstract, with the expectation that the student will supply an effective method in order to solve the problem. In the absence of specific prior knowledge, problem solvers must search for solutions by means-end analysis or trial and error (Schnotz and Kürschner, 2007). As noted, this makes high demands on working memory as the student has to search their long-term

³⁵ This definition is now very familiar to teachers and school leaders in English schools, not least because it has been taken up by Ofsted: Extract from School Inspection Handbook, 2022: ‘222. Inspectors will be alert to unnecessary or excessive attempts to simply prompt pupils to learn glossaries or long lists of disconnected facts. Learning can be defined as an alteration in long-term memory. If nothing has altered in long-term memory, nothing has been learned. However, pupils learn by connecting new knowledge with existing knowledge. Pupils also need to develop fluency and unconsciously apply their knowledge as skills. This must not be reduced to, or confused with, simply memorising disconnected facts. When inspectors evaluate the impact of the education provided by the school, their focus will primarily be on what pupils have learned.’ <https://www.gov.uk/government/publications/school-inspection-handbook-eif/school-inspection-handbook> accessed 11.07.2023

memory for appropriate methods and then retrieve and test them. An example of such an abstract problem is shown in Figure 7:

Figure 7 - Abstract Mathematics Problem

7 (c) Solve $2x - 1 > 9$

.....

.....

Answer (2 marks)

Source: AQA GCSE Mathematics Unit 02: Number and Algebra (Higher tier), November 2013.³⁶

In the problem shown in Figure 7, no assistance or method is given; students are instead simply directed to solve the problem. They will need to bring their own methods and understanding to the question. This demand will pose a considerable cognitive load on the students.

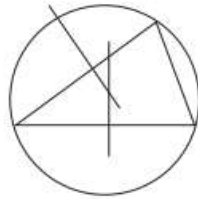
The second approach, studied by Sweller and others in the late 1980s, is to create goal-free problems, such as worked-out examples (Sweller and Levine, 1982; Sweller and Cooper, 1985), which give structure to the problem-solving approach, and lead to a task for which the method has already been given or hinted at.

An example of a problem adopting this second approach is given in Figure 8. In this question, cognitive load is reduced: the problem is less abstract, and a step-by-step method is specified – in fact, it might be argued that, in this particular case, there is no problem for the student to solve, but rather a set of instructions to follow. It would be helpful for the student to know beforehand what is meant by a ‘perpendicular bisector,’ but even this is not necessary prior knowledge, because the small diagram helps the student to guess or to contextualise the term.

³⁶ This question and the two that follow are taken from a paper set in 2013. Analysis of papers from 2013 to 2021 shows that these styles of questions have featured consistently in Mathematics GCSE papers during all the years of this research study.

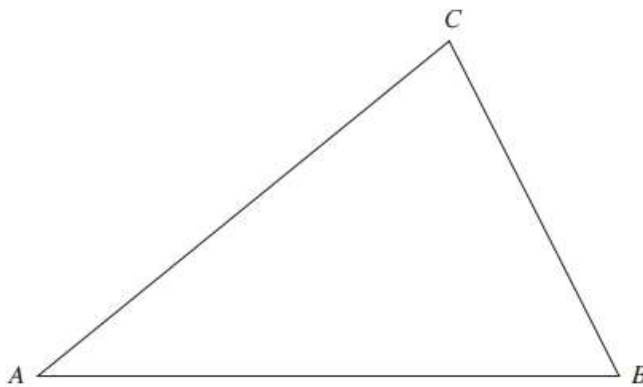
Figure 8 - Mathematics Problem with a Structured Method

13 Use these steps to construct a circle passing through the vertices of the triangle ABC .



- Construct the perpendicular bisector of AB .
- Construct the perpendicular bisector of AC .
- Use the point of intersection of the bisectors as the centre of the circle.
- Draw the circle through A , B and C .

Show your construction arcs clearly.



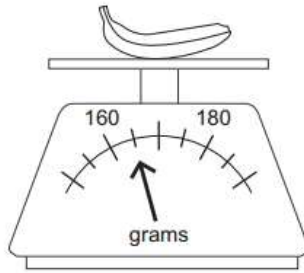
(4 marks)

Source: AQA GCSE Mathematics Unit 03 (Higher tier), November 2013.

It is also possible to hint at a method without making it so explicit or proscriptive. In the next example, Figure 9, the learner is led through the problem one step at a time. The method, although not explicitly stated, is inferred.

Figure 9 - Mathematics Problem, with Step-by-step Approach

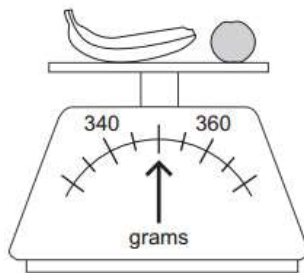
5 A banana is weighed.



5 (a) How much does the banana weigh?

Answer grams (1 mark)

5 (b) The banana is now weighed with an orange.



How much does the orange weigh?

.....
.....

Answer grams (2 marks)

Source: AQA GCSE Mathematics Unit 03 (Foundation tier), November 2013.

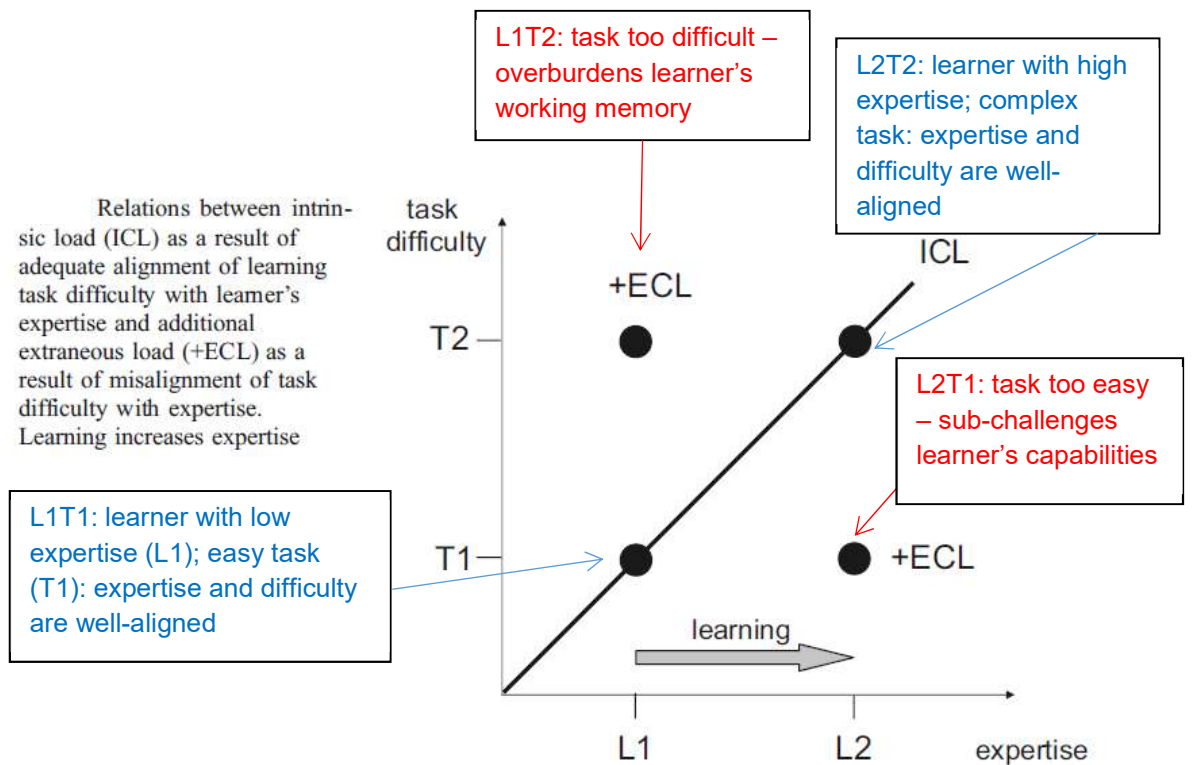
The step-by-step approach in the question within Figure 9 imposes a slightly higher cognitive load than the question in Figure 8, because the student has to construct a method, and work out that the mass (let this be represented by 'm' of the orange 'o') is equal to the total mass of the 'orange plus banana' minus the mass of the banana ('b'). In mathematical terms, this can be stated as,

$$m(o) = m(o+b) - m(b).$$

The step-by-step method imposes a considerably lower cognitive load on the student than the plain instruction 'solve' in the higher level algebra question (Figure 7). This is because the method is hinted at, and information is given in a pictorial form that is uncluttered by additional labelling; thereby avoiding the redundancy effect: information is, as a consequence, relatively easy to assimilate.

Schnotz and Kürschner (2007) looked in more detail at the relationship that exists between intrinsic cognitive load and a learner's ability for a task. In Figure 10, points that lie below the intrinsic cognitive load (ICL) diagonal line show that task difficulty is below the level of learner expertise (that is, it is too easy), whereas points above the ICL line show that task difficulty exceeds the level of learner expertise (the task is too difficult). This led Schnotz and Kürschner to assert that, 'instruction has to align learning task difficulty with the learner's level of expertise' (2007, p. 485). This is why, as presented in the model of the Utopian examination system in the introductory chapter of this study, examination papers have questions with different levels of demand – usually, steadily increasing throughout each examination paper – so that the specific level of expertise of each student can be ascertained, by the number and complexity of the questions they answer correctly. Subjects including mathematics in the English GCSE system have two tiers of entry, higher and foundation, so that the difficulty of test items can be better aligned with students' level of expertise. He *et al.* comment that, even so, 'use of examination time is inefficient, as pupils are expected to answer questions which are either too difficult (for less able pupils) or too easy (for more able pupils)... Both low ability and high ability pupils may have a demoralising and demotivating experience' (2015, pp. 80, 82. Note that He *et al.* do not attempt to define what they understand by "able" or "ability" in this context).

Figure 10 - Relationship between Intrinsic Load and a Learner's Expertise



Source: Adapted from Schnotz and Kürschner, (2007, p. 478); Boxed labels: author's own.

Task demand, which, as was suggested earlier, equates to intrinsic cognitive load (the diagonal line ICL, in Figure 10), may be reduced using worked examples, as was seen in the mathematics question in Figure 8. Presentational features of examination questions may, however, increase cognitive load. Moreover, they may do so in non-desirable ways. Such features may include split-attention effects, where the student needs to look at a diagram and at instructions that are printed separately from the diagram; and redundancy effects, where unnecessary additional information is included which may cause the student to have to discriminate, identify, and filter out the redundant information (Sweller *et al.*, 1998). Extraneous cognitive load is affected by the interaction that occurs between relevant information and the working memory; interactivity with irrelevant information represents a waste of the student's time and mental effort.

3.1.3 Optimising cognitive load, and Vygotsky's Zone of Proximal Development

Although Sweller and his research partners repeatedly asserted that, in general, the goals of instructional design were to optimise intrinsic cognitive load and reduce extraneous cognitive load, the reduction of cognitive load is not always helpful for learning. Since GCSE examination questions are often used in a classroom context, within a sequence of teaching and learning, it is highly relevant to consider this effect here. Schnotz and Kürschner (2007) explained this apparent paradox by reference to the work of the Soviet-era psychologist Lev Vygotsky,³⁷ and in particular his theory of the Zone of Proximal Development. The zone of proximal development has been defined as,

‘The distance between the actual developmental level as determined by independent problem solving and the level of potential development as determined through problem-solving under adult guidance, or in collaboration with more capable peers’ (Vygotsky, 1978, p. 86).

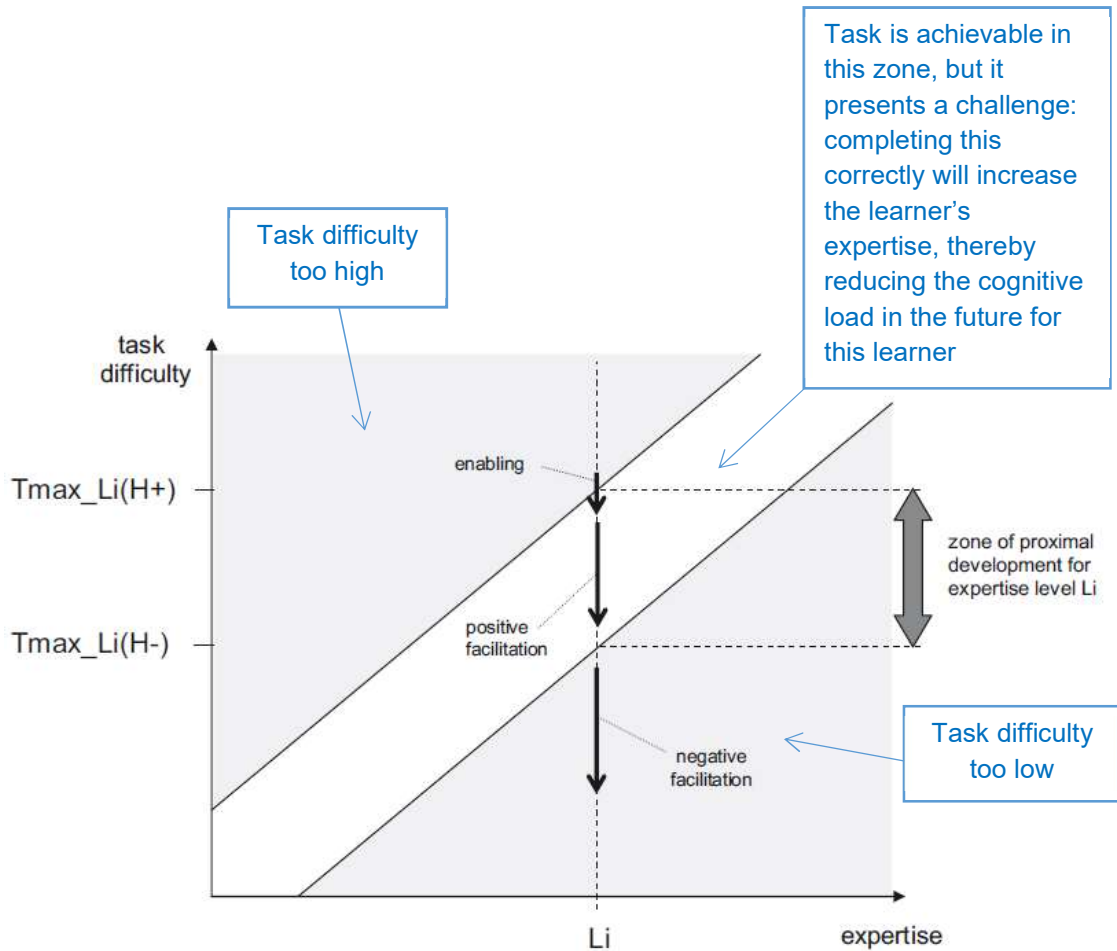
An essential feature of Vygotsky's theory is that, in order for a student to develop in expertise, a ‘more knowledgeable other’ is needed: a teacher, for example (Vygotsky, 1978, p. 86); social interactions must occur between the teacher and the student; and supportive activities are needed.

Figure 11 shows the relative difficulty of a task in respect of a student of a particular level of expertise. Vygotsky's premise was that learning happens best when a task is “stretching” for the student: it is possible but not too easy. Figure 11 is essentially the same diagram as Figure 10, except that the diagonal intrinsic cognitive load line from Figure 10 has broadened into a “zone” in which learning takes place. Vygotsky (1978, pp. 86-88) explained that ‘Positive facilitation’ is intervention by the ‘more knowledgeable other’ that enables learning to take place: the difficulty of the task, in effect, is lessened by the supportive activities that are presented. ‘Negative facilitation,’ on the other hand, is intervention that makes the task too

³⁷ Lev Vygotsky (1896-1934), psychologist and social theorist. Vygotsky was born in Belarus and worked in Moscow. His theories did not circulate widely in western educational circles until after his death, when his writings were translated into English and published in the United States of America in the 1960s and 1970s.

easy: the student is now able to perform the task, but since it no longer brings sufficient challenge, it has an instructionally negative effect, and the student does not learn (Schnotz and Kürschner, 2007, p. 486).

Figure 11 - Vygotsky's Zone of Proximal Development



Source: Adapted from Schnotz and Kürschner, (2007, p. 486); Boxed labels: author's own.

According to Vygotsky, teaching activities that aim to promote learning should include learning tasks within the zone of proximal development. Schnotz and Kürschner explained that:

- 'If task difficulty is higher than the zone of proximal development,
 - Student's cognitive capacity is overwhelmed, because total cognitive load exceeds student's working memory capacity;

- Therefore the probability of success is 0%; that is, the task is too demanding to be accomplished by this student with this level of expertise.
- If task difficulty is below the zone of proximal development,
 - Student is sub-challenged,
 - Therefore the probability of success is 100%; that is, the task is too easy for the student' (Schnotz and Kürschner, 2007, pp. 486-7).

The practical application of Vygotsky's theory, and its relevance to this study, is that when teacher sets practice questions for a student, the demands of the questions should be located within the student's zone of proximal development so that the student is challenged slightly, and their expertise increased by their being required to rise to the challenge and succeeding. Where the total cognitive load of a task exceeds the student's working memory capacity, the task is too difficult; where the student's expertise is such that they have surplus working memory capacity, they will find the task easier. It is therefore also possible that a student may extend their learning in an examination context when they encounter a question the demands of which are located within their zone of proximal development. In practice, this may be a rarer occurrence.

Schnotz and Kürschner (2007) showed how cognitive load theory can be applied to test design; this can be extended to the specific case of GCSE examinations. For teachers applying cognitive load theory, the focus is to assist instructional design, so that students can be helped to develop their expertise. For test designers (in this study, examiners), the approach is to attempt to measure students' expertise at a given time: the examiners' goal is that students' test scores can be translated directly and reliably into a reading of their expertise or ability. The task under discussion in cognitive load theory becomes the test item or question in the examination, and imposes a certain intrinsic cognitive load (demand); how well a student is able to withstand that cognitive load depends on their level of expertise. In cognitive load theory terms, this is a measure of how well the given student is able to manage their working

memory capacity, and how effective the learning process has been in transferring problem-solving capacity from working memory to long-term memory.

The goal of the test designer/examiner, then, is to create a range of questions that can give an accurate reading of the expertise of the student by imposing a task of a certain (previously determined) cognitive load. These will be 'exams that make you think' (*Guardian*, 2013), or, in the title of Heller-Sahlgren's 2014 book, 'Tests worth teaching to'. Typically, there will be items beyond which the less expert will not pass, as well as items that will tax the working memory capacity of even the most expert.

For the purposes of assessment design, it needs to be assumed that a student's expertise can be measured, meaning that it is fixed, at least for the duration of the test/examination. The test result is a "snapshot in time" of the student's expertise. There is a potential threat to validity here, however. Examination questions are put into the public domain once the examination has been taken. Teachers use past examination questions for instruction, and students use them for learning and revision, with the explicit intention that exposure to examination questions (within the zone of proximal development, at first under the guidance of the 'more knowledgeable other' and gradually more independently) will enable students to apply, stretch, and thereby increase their expertise. Examiners, however, wish to create a test that is reliable, that is, the same grade would be achieved if the test were to be repeated (although, in practice, questions are rarely repeated). These conflicting positions can be reconciled only if it is assumed that a student's expertise increases sufficiently slowly so as not to be a threat to the reliability of the examination. The student never takes the same examination a second time, but reliability remains a contested concept within the public examination system (see, for instance, Baird and Black, 2013).

In practice, for a small proportion of students, however, the contrary position might be argued: that learning takes place more quickly under conditions of stress, such as those imposed by examinations and tests. In these circumstances, since learning increases expertise

(which, in the definition cited earlier, has the effect of reducing intrinsic cognitive load), the balance between intrinsic cognitive load and expertise is *not* fixed, even in the short term, for an individual student. If a student were to succeed in solving problems under examination conditions that they had not been able to solve previously, the expertise of this student could be described as having increased rapidly.

3.1.4 Critical engagement with Cognitive Load Theory

Cognitive load theory has proved to be an adaptable theoretical framework that has absorbed criticisms (for instance, Moreno, 2006, in application to worked examples; Schnotz and Kürschner, 2007, around conceptual clarity) and used them to strengthen its model. Its key insights and contributions to instructional design have been the suggestions that, to maximise learning effectiveness, it is helpful to:

- present material that aligns with the prior knowledge of the learner;
- avoid non-essential and confusing information;
- stimulate processes that lead to conceptually rich and deep knowledge.

Critically engaging with cognitive load theory, it is possible to observe that the conceptual framework has limitations, and that some of its assertions and insights are problematic, particularly in the context of this current study. Whilst making a number of evaluative criticisms of cognitive load theory, De Jong acknowledges that ‘an important point to note is that cognitive load theory is constructed in such a way that it is hard or even impossible to falsify’ (2010, p. 125). Nonetheless, a number of concerns have been raised by authors and commentators; these are principally conceptual, practical, and in terms of the implications for teaching critical thinking.

De Jong (2010) raises a number of questions around cognitive load theory that can be related directly to this current study, among which the most significant concerns are whether the different types of cognitive load can be distinguished in practice, whether cognitive load can

be measured in any meaningful and consistent way, and the lack of clarity around how cognitive load is actually related to instructional design.

Cognitive load theory distinguishes between intrinsic and extraneous cognitive load, and asserts that intrinsic cognitive load is fixed and cannot be changed by instructional treatments (Ayres, 2006; Sweller *et al.*, 1998; Paas *et al.*, 2003). However, Van Merriënboer *et al.* (2003) showed that sequencing instruction in a simple-to-complex order enabled the control of intrinsic cognitive load, so that learners did not experience the full complexity at the start. Chi (1992, p. 2005) demonstrated that students who developed ontological misconceptions (such as seeing “force” as a material substance) found them very hard to change subsequently, which made the learning of scientific concepts even harder to understand. These insights have important implications for teachers, who could effectively reduce intrinsic cognitive load by the use of careful instructional design to avoid misconceptions and sequence learning carefully.

The measurement of cognitive load is problematic. Although some empirical measures have been attempted, it is not clear what these actually mean in practice, since it is necessary to regard cognitive load as being always relative. Attempts to measure cognitive load (Paas *et al.*, 1992; Mayer *et al.*, 2005; Paas *et al.*, 2003) have relied on self-report questionnaires typically completed by students after they have finished an episode of learning. There is, however, no standard form to the questions used, and reports may vary considerably depending on the wording and specific details of the questions asked. Outcomes are therefore inconsistent, and there is no agreed-upon meaning of the findings of the reports. In this study, therefore, no attempt has been made to *quantify* the cognitive load experienced by students in approaching examination questions, but rather to understand the sources and nature of the load.

The distinction between intrinsic and extraneous cognitive load may not always be clear-cut, particularly so in the case of examination questions: in fact, it depends on the interpretation of what the *cognitive task* of the question is. To give an example, a mathematical question may

contain a number of pieces of information, not all of which are needed for the student to use in their answer; the question may be posed in the form of an extended paragraph, or it may relate to a real-life context. One student, eager to get to grips with the actual mathematics in the question, and thinking that the *task* of the question is simply to manipulate the numbers and create an answer, may be frustrated with this presentation. To use cognitive load theory terminology, they may experience all of the redundant information, the additional wording and the context, as sources of extraneous cognitive load, creating distraction and redundancy effects. On the other hand, a second student may perceive that the question is actually testing their problem-solving skills, and that the first *task* they need perform is to filter through the information or the paragraph to decide what is relevant, before constructing their answer to the question. What appeared as extraneous cognitive load to the first student is understood by the second student as part of the problem to be solved, that is, it is part of the intrinsic cognitive load. To test this idea further in the current study, discussions will be held with students around their understanding and interpretation of examination questions, and how they approach more complex questions.

De Jong (2010) also notes that studies and measurements of cognitive load do not relate to the amount of time given for a learning activity. This does seem unrealistic: a problem that feels hard to solve in two minutes might appear more straightforward if 20 minutes were available, yet the intrinsic cognitive load is presumably the same. In terms of cognitive load theory, the implication is unclear – does time pressure constitute part of extraneous load? In the context of examinations, this could be particularly important, since a time constraint in an examination might radically alter the student's perception of the difficulty of a question, and their ability to meet its demands. The effect of stress on students' performance is a facet of examinations that will be explored briefly in this study.

In his response to cognitive load theory, Ellerton (2022) contrasts two approaches to critical thinking. On the one hand is an approach that treats critical thinking almost as a curriculum in

itself, consisting of a range of skills, knowledge and dispositions that can be taught. On the other hand, an approach taken by proponents of cognitive load theory may be summarised as the idea that critical thinking cannot be discussed meaningfully outside of the context of a subject discipline, and that it can be developed only through deep engagement with subject knowledge. Ellerton proposes that this is a false dichotomy, and that 'learning to think and thinking to learn' (Ritchhart and Perkins, 2005, p. 795) can be complementary and linked rather than separate or even mutually exclusive. Criticisms of cognitive load theory articulated by Ellerton focus more on practical than conceptual concerns. These include definitional problems (what *is* critical thinking?); and the impossibility of disproving either dichotomous position experimentally. In terms of practical teaching, Ellerton observes that cognitive load theory's focus on a tight but generalised view of what makes instructional design does not move easily to contexts applicable all over the curriculum (in subjects such as history and English, music and drama, physical education and technology, for example); and that it does not take into account an appreciation of the learner as an autonomous agent. Differences in students' character and disposition, he argues, might have profound effects on the effectiveness of instructional approaches informed by cognitive load theory (or any other pedagogical approach, it might be added). In other words, whilst cognitive load theory might outline how an approach informed by its insights might work best in optimising instructional design for a generalised situation, in practice its effectiveness may vary considerably with different learners. (Even when all the features of cognitive load theory are optimised, there is no guarantee that learning will take place.) Ellerton suggests that it is critically important that cognitive load theory be applied to a range of real-world applications, in order to investigate and refine it further. This current study makes a small contribution to this next step, not testing cognitive load theory in practice, but bringing it to bear on the insights that students have into their own learning experiences.

In conclusion, cognitive load theory conceptualises and theorises the demands that students face as they learn any new material. It is grounded in evolutionary psychology, and it links

closely to theories about working memory capacity. Cognitive load theory states that the total cognitive load a student encounters is the combination of intrinsic and extraneous load. In this study, the cognitive processes for students facing an examination question are considered to be equivalent to those operating in the classroom. Insights gained from cognitive load theory can therefore be applied to examination questions. The core demands of examination questions can be equated to intrinsic cognitive load, and anything else in the examination question that distracts or confuses students can be identified as contributing to extraneous cognitive load. Extraneous cognitive load may produce unwanted or unexpected sources of difficulty, as experienced by students. In terms of GCSE examination questions, if examiners are to produce results that are as reliable a reading as possible of students' abilities and expertise, their questions need to reduce extraneous load as much as possible, so that students can focus on the intrinsic load of the topic and the problem that is being tested.

3.2 Taxonomies of learning, student responses and question demands

In order to understand the difficulties faced by students in examinations, and to investigate and discuss the extent to which students themselves understand these difficulties, it is necessary to use a shared vocabulary and classification for their understanding. In this section of the literature review, attention is therefore turned to the different ways in which educational thinkers have sought to understand the processes and outputs of learning. In so doing it focuses on the models put forward by Bloom *et al.* (1956, revised 2001), Marzano (2001 and 2007), and Burch (Four Stages of Competence model, 1970s). This section also surveys the CRAS scale for assessing the demands of examination questions (Pollitt *et al.*, 2007). Students' responses and the understanding they communicate are subsequently discussed in Chapters 5, 6, and 7. Bloom's Taxonomy is frequently encountered in conversations with teachers, and this is the starting point for this review.

3.2.1 Bloom's Taxonomy

Following discussions at the 1948 Convention of the American Psychological Association, Benjamin Bloom led a group of educators in the United States of America to classify educational goals and objectives. Taxonomies (classifications) were planned of three domains:

- Cognitive domain: knowledge-based, consisting of 6 levels
- Affective domain: attitudinal-based, 5 levels
- Psychomotor domain: skills-based, 6 levels

Bloom and his colleagues completed only the first classification, the cognitive domain. This was published in 1956 as *Taxonomy of Educational Objectives* (Bloom *et al.*, 1956). As Forehand has explained, this taxonomy was a 'multi-layered model of classifying thinking according to six cognitive levels of complexity' (2010, p. 2). Whilst Bloom's Taxonomy has become extremely influential in English-speaking educational circles, its ubiquity has sometimes come at the price

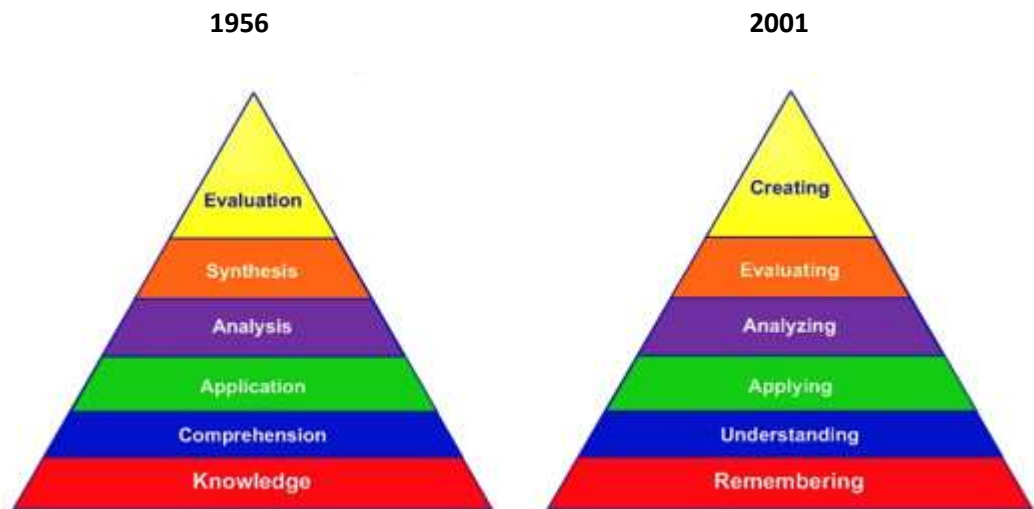
of oversimplification and a loss of understanding of the objectives (and limitations) of the original approach.

The fundamental concept underlying Bloom's Taxonomy is that it is possible to classify or separate out the thinking activities associated with learning, and that these activities can be understood as hierarchical. Bloom *et al.*, set these in a strict hierarchy, often since represented as a pyramid (although, as Stern, 2018, notes, this pyramidal representation appears nowhere in the original 1956 work), resting on a base of knowledge. It is the role and function of knowledge that has been most contentious.

At the time of both the original discussions (1948) and the publication by Bloom *et al.*, (1956), the physiological understanding of thinking ("brain science") was in its infancy. Until the invention of magnetic resonance imaging (MRI) by Raymond Damadian in the 1970s, it was possible to deduce the associations between cognitive functions and the activity of particular parts of the brain largely only through case studies of patients who had survived catastrophic brain injury. MRI scans, being non-invasive, offered great potential to inform studies of brain function. This capability was developed from the 1980s onwards, and resulted in a radical rethinking of how humans learn and where in the brain different forms of thinking take place.

In 2001, Anderson and Krathwohl (former student and colleague respectively of Bloom), working with others, published a revision of Bloom's *Taxonomy* (Anderson *et al.*, 2001). At first sight, the differences may appear slight, moving from nouns to verbs ("comprehension" becomes "understanding", for example), and transposing the two upper layers of the pyramid (see Figure 12). Their revision did, however, go much further than this but much of their commentary has been overlooked. The lack of engagement with the commentary of Anderson, *et al.*, by many teacher educators, has caused Bloom's Taxonomy to be oversimplified and, arguably, misapplied, diminishing its potential impact. In other ways, as Marzano and Kendall (2007), and Kagan (2005) have pointed out, inherent structural weaknesses were retained in the 2001 revision by Anderson, *et al.*

Figure 12 - Bloom's Taxonomy: 1956 Original and Anderson et al.'s 2001 Revision



Source: Berger (2018).

One of the most noticeable features of – and the most significant problems with – Bloom’s model is the ascending nature of the cognitive skills: the striped horizontal layers seen in Figure 12. As Agarwal explained:

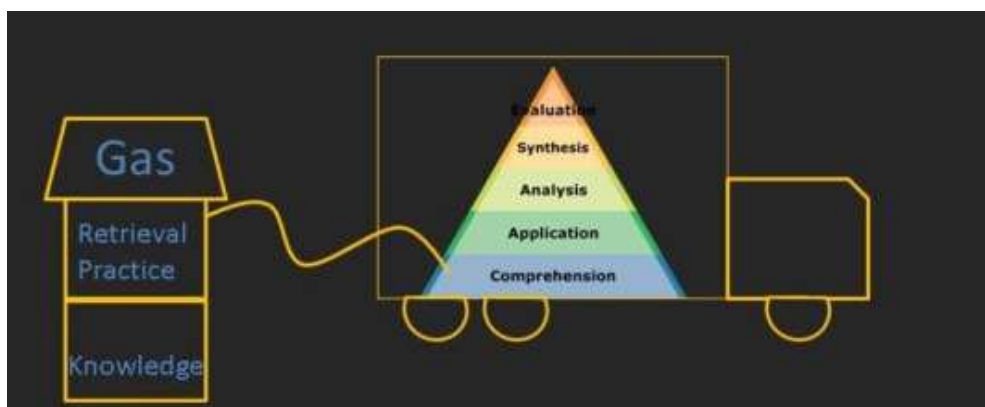
‘In part because of its simplicity, Bloom’s taxonomy has contributed to the collective notion that foundational knowledge (literally the foundation or base of the pyramid) precedes higher order learning (the categories located higher in the pyramid)’ (2019, p. 190).

From this, a tendency has developed which labels the lowest three forms of cognition (knowledge, comprehension, and application) as “lower-order thinking skills” and the upper three (analysis, synthesis and evaluation) as “higher-order thinking skills.” This categorisation was, in part, a response to the concerns of American teachers in the 1980s and 1990s, following the publication of state-wide test results in the United States in the 1970s, which suggested that too much teaching was merely ‘instructional’ (fact-based) and that students were not being taught or expected to apply their knowledge and answer more evaluative questions (Marzano and Kendall, 2007). The pedagogical pendulum then swung in the other direction, leading teachers and other educators to dismiss these “lower” cognitive functions, as Lemov described:

‘Generally when teachers talk about “Bloom’s taxonomy,” they talk with disdain about “lower level” questions. They believe, perhaps because of the pyramid image which puts knowledge at the bottom, that knowledge-based questions, especially via recall and retrieval practice, are the least productive thing they could be doing in class. No one wants to be the rube at the bottom of the pyramid’ (Lemov, 2017, online)

Lemov offered a corrective (Figure 13), and asserted that knowledge is the essential ‘fuel that allows the engine of thinking to run,’ supplying all the other cognitive functions.

Figure 13 - Lemov: *'Bloom's Delivery Service'*



Source: Lemov (2017, online)

In Lemov’s graphic (Figure 13), knowledge “fuels” all the other cognitive skills. More recently, Berger has argued that the 2001 revision of Bloom’s hierarchy did not go far enough:

‘The problem is that both versions present a false vision of learning. Learning is not a hierarchy or a linear process. This [pyramid] graphic gives the mistaken impression that these cognitive processes are discrete, that it’s possible to perform one of these skills separately from others. It also gives the mistaken impression that some of these skills are more difficult and more important than others. It can blind us to the integrated process that actually takes place in students’ minds as they learn’ (2018, online).

Berger argued for a more fluid understanding of learning in which knowledge and cognition are essential, but where they are neither divorced from, nor merely foundational to, the acquisition of other learning skills:

‘The root problem with the framework is that it does not accurately represent the way that we learn things. We don’t start by remembering things, then understand them, then apply them, and move up the pyramid in steps as our capacity grows. Instead,

much of the time we build understanding *by applying knowledge and by creating things*' (Berger, 2018, online, emphasis in original).

In 1994, Furst recognised the essential weakness in the original classification, namely, its assumption that cognitive processes are ordered on a single dimension of simple-to-complex behaviour. Subsequently, Anderson, *et al.*, (2001) also came to the belief that knowledge was both fundamental to and inextricably bound to other aspects of cognition. In addition to the semantic differences in their 2001 revision, therefore, they also proposed a change of structure. Bloom's 1956 version is one-dimensional; in its revised form, Bloom's is a two-dimensional table, (as shown in Figure 14), which plots the knowledge dimension on the vertical axis (the kind of knowledge to learn) against the cognitive process dimension (the process used to learn) on the horizontal axis.

Figure 14 - Bloom's Taxonomy (Revised): The Knowledge Dimension

Knowledge Dimension	Cognitive Process Dimension					
	Remember	Understand	Apply	Analyze	Evaluate	Create
A. Factual knowledge						
B. Conceptual knowledge						
C. Procedural knowledge						
D. Metacognitive knowledge						

Source: Adapted from Krathwohl (2002, p. 216)

Their reworking of the knowledge dimension showed that Anderson, *et al.*, (2001) recognised the linear shortcomings of Bloom's original model. Their revision asserted that knowledge is such an important part of learning that its function can hardly be overstated: it is not simply a threshold over which the learner passes *en route* to more sophisticated forms of learning, but it is fundamentally linked to all "higher" forms of learning and cognition. Through this lens, it

can be seen that there can be value in examination questions that test cognitive skills at different levels: factual recall and understanding, as well as application of knowledge, interpretation of data, and evaluative skills. This levelled analysis had an attractive simplicity, but this was also its major flaw, and insights available from psychology and brain science exposed this.

The idea that knowledge links to all other cognitive skills, and that extra dimensions of complexity exist within each layer, was taken further by Kagan (2005), who offered a more fundamental critique of Bloom's hierarchical taxonomy. Supported by examples from brain science, he disagreed that Bloom's model conveyed anything of real meaning about the complexity of different thinking skills. As Kagan noted:

'At first glance it makes sense to think of recall as less complex than evaluation. It feels like we recall effortlessly (memories just pop to mind), whereas evaluation takes concentration, and a good evaluation involves careful weighing an [of] outcome against one or more criteria. Upon reflection, however, we discover any of the thinking skills can be very simple or very complex depending on how deeply we engage that particular type of thinking. If I ask you if you like chocolate ice cream (an evaluation level question), the answer simply pops to mind as immediately and effortlessly as if I ask you if you ate chocolate ice cream within the last hour—a recall level question. If I ask you to recall all the times in the last month you ate or saw ice cream, the answer demands a great deal of cognitive effort, just as if I asked you to evaluate all the pros and cons of eating ice cream. Evaluation, recall, and any other thinking skill can be engaged at a simple or complex level. *Complexity is not associated with the type of thinking skill, but rather with the level at which the thinking skill is engaged*' (Kagan, 2005, online, emphasis in original).

In this quotation, Kagan points out how an evaluative ("higher order" thinking task can be straightforward, and a recall ("lower order") task can be complex and demanding, the exact opposite of their positions in Bloom's hierarchy. Kagan asserted that complexity – which, as is subsequently discussed, can make up part of the demand or cognitive load imposed – is determined by the level rather than the type of thinking. This view therefore challenged the ability of Bloom's framework – whether original or revised – to adequately describe the real experiences of thinking and learning.

Bloom's taxonomy has, nonetheless, exerted considerable influence on the construction of examination specifications. Pollitt, Ahmed, and Crisp (2007) observed that,

'Since the introduction of the O level and O grade examinations³⁸ it has been standard practice to specify the content of papers in terms of cognitive skills or 'assessment objectives' (AOs). These have generally been derived from the taxonomy of cognitive 'objectives' for education of Bloom (1956), except in the cases of languages, art, and so on' (2007, p. 182)

However, the use of Bloom's taxonomy has been implicit rather than systematic, and there has been little rigorous evaluation: 'there are very few studies, and no significant comparability studies, where judges have been asked to classify individual questions in terms of Bloom's taxonomy' (Pollitt *et al.*, 2007, p. 182). Similarly, McLone, and Patrick noted that skilled examiners are

'Able to recognise 'demand' and generally to agree in estimating the overall level of demand in questions. However, they were much less good at explaining it; they could not analyse a question to describe the cognitive elements and processes that were the source of that difficulty' (McLone and Patrick, 1990, cited in Pollitt *et al.*, 2007, p. 184).

With regards to this study, the importance of the general shape and principles of Bloom's Taxonomy is two-fold. First, they have asserted an influence on the construction of examination specifications. Secondly, individual examiners and examination boards appear not to have engaged with Bloom's Taxonomy in any systematic way.

In spite of the rather approximate way in which Bloom's Taxonomy has often been understood and applied, its ubiquity and dominance in educational thinking and training in the English-speaking world continues³⁹. This is not matched, however, by a similar volume or richness of evaluative study focused on educational practice with students. In the course of this literature search, using multiple searches and Boolean operators on university library search engines,

³⁸ 'O' level examinations, taken by 16-year-old students, were introduced in 1951. They were replaced by GCSEs in 1988.

³⁹ Not entirely without criticism: see, for example, Case (2013, p. 196): 'in addition to enduring popularity, it is arguably one of the most destructive theories in education'; but even many of its detractors do not question its hierarchical assumptions. Case, for instance, implicitly accepts these when he writes that teachers should 'adjust the difficulty so that every student engages regularly in "higher order" learning activities.'

Google Scholar, and the Institute of Education Sciences (eric.ed.gov), up to the end of 2021, more than 100,000 studies, books, articles and references were found that implement, apply, and use Bloom's taxonomy to teaching at different levels and in different subjects. But only three were discovered that sought to evaluate whether teaching approaches based on Bloom's taxonomy demonstrated actual efficacy in studies with students. These three papers all analysed studies with undergraduate or postgraduate university students (Chan *et al.*, 2002: 17 postgraduate social work students; Crowe *et al.*, 2008: 3 small studies, a total of 178 undergraduate students; Agarwal, 2019: 48 undergraduate students); not a single evaluative study was discovered that involved school-age students. This dearth within existent literature once more underlines the research gap that this thesis addresses.

3.2.2 Marzano and Kendall: A New Taxonomy

Like Kagan, Marzano and Kendall (2007) observed that a critical problem with Bloom's taxonomy and any attempted revision⁴⁰ was that 'it attempted to use degrees of difficulty as the basis for the differences between levels of the taxonomy' (p. 11). Marzano and Kendall judged that:

'Ultimately, any attempt to design a taxonomy based on difficulty of mental processing is doomed to failure, because of the well-established principle in psychology that even the most complex of processes can be learned at the level at which it is performed with little or no conscious effort' (2007, p. 11).

They gave the example of driving a car fast on a busy road, a task which might seem impossible to a novice driver but which, after sufficient experience, many people can accomplish without conscious effort, to the extent that they can also simultaneously undertake other tasks (such

⁴⁰ Marzano and Kendall judged that the revision of Bloom's taxonomy by Anderson, *et al.*, 'suffers from the same inherent weakness... the tacit assumption that its levels are ordered hierarchically in terms of difficulty' (Marzano and Kendall, 2007, p. 17).

as conversation). Instead of relying on a fixed idea of difficulty of any specific cognitive skill, they described the difficulty of a mental process as:

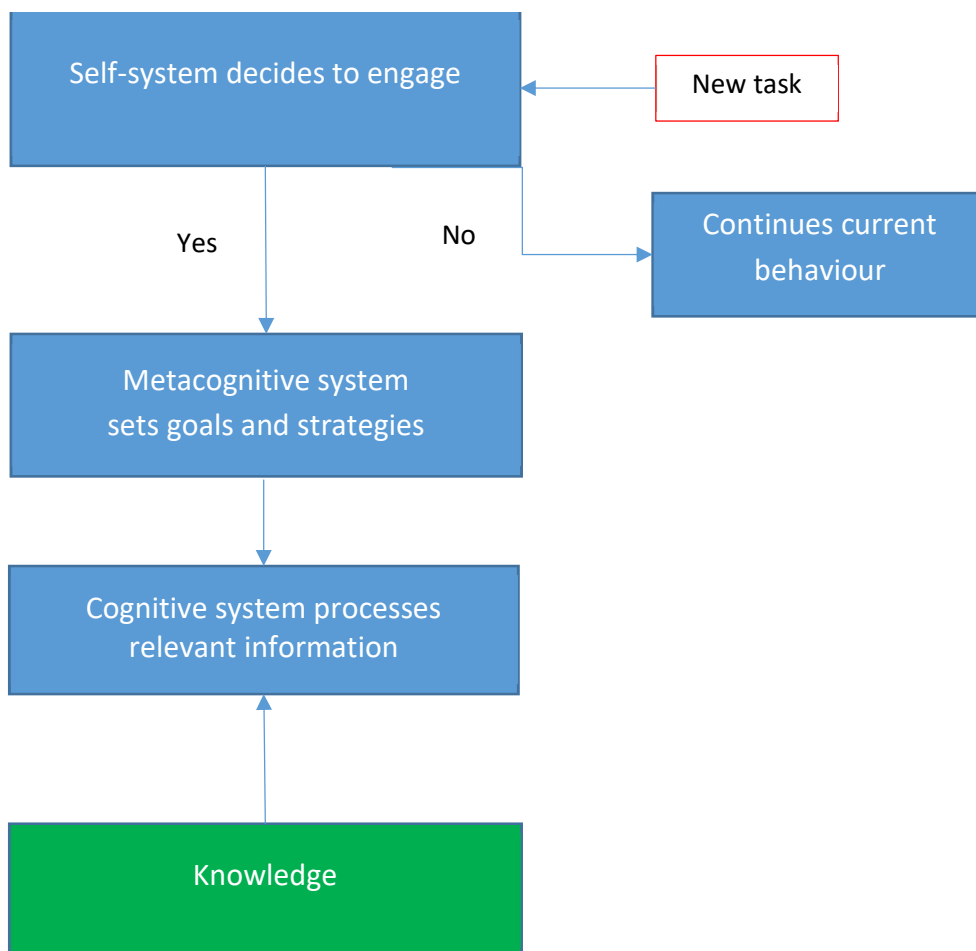
‘A function of at least two factors – the inherent complexity of the process in terms of steps involved and the level of familiarity one has with the process’ (Marzano and Kendall, 2007, p. 11; emphasis added).

Marzano and Kendall observed that, ‘although mental processes cannot be ordered hierarchically in terms of difficulty, they can be ordered in terms of control: some processes exercise control over the operation of other processes’ (2007, p. 11). This is the organising concept of their ‘New Taxonomy,’ as represented in Figure 15. The relevance to the present study is that answering an examination question necessarily involves a student in operating various different mental processes. The extent to which Marzano’s and Kendall’s description of a mental process resonates with students’ experiences of answering examination questions is explored in Chapters 5 and 6.

Marzano and Kendall categorised the cognitive system into four ‘domains of knowledge: knowledge retrieval, comprehension, analysis, and knowledge utilization’ (2007, 11). In this respect, their model shares some features with Bloom’s taxonomy, except that each one of these four domains of knowledge can be engaged in ‘different levels of processing’, creating a two-dimensional model. Above this cognitive system sits ‘the metacognitive system’ and, above this, ‘the self-system’ (Marzano and Kendall, 2007, p. 11). At the highest level within Marzano’s and Kendall’s model, the self-system makes a decision about whether to engage with the new information the student receives – in examinations, and in the present study, this new information is the examination question. The student’s self-system ‘contains a network of beliefs and goals’: it not only makes a decision about whether to engage but is ‘also a prime determiner in the motivation one brings to a task’ (2007, p. 12). The self-system, then, is the ‘home’ of self-efficacy, the belief that one can accomplish a task. The metacognitive system ‘sets goals relative to the new task’; it also ‘designs strategies for accomplishing a given goal

once it has been set' (Marzano and Kendall, 2007, p. 12). The metacognitive system, once engaged, continually interacts with, and regulates, the cognitive system.

Figure 15 - Marzano and Kendall: Model of Behaviour

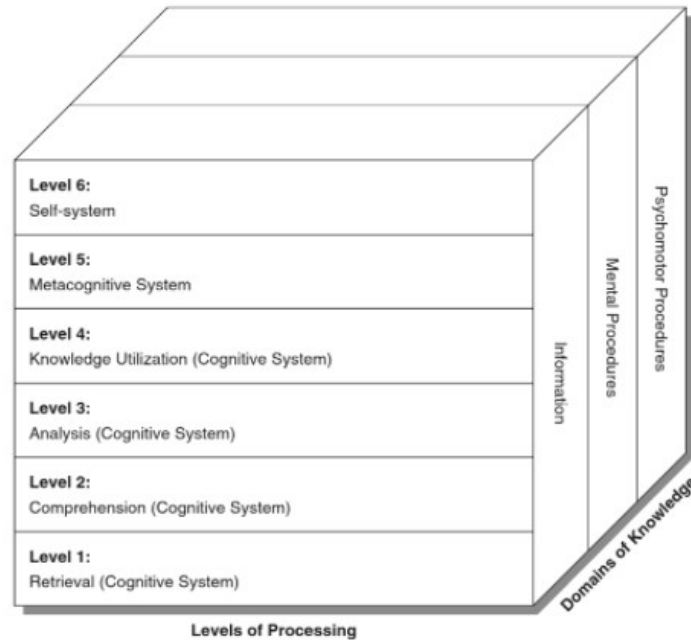


Source: Marzano and Kendall (2007, p. 11).

Marzano and Kendall further divided the cognitive system into four levels: retrieval, comprehension, analysis, and knowledge utilization (see Figure 16). This stratification may appear superficially similar to Bloom's taxonomy, and it has led to criticisms that it, too, is a linear or unidimensional model (see, amongst others, Irvine, 2017, and Greatorex *et al.*, 2019). Crucially, however, unlike Bloom, Marzano and Kendall threaded "information" (that is, knowledge) through all four levels of the cognitive system, and they explained how the

metacognitive system reviews, regulates, and deploys the four different levels of the cognitive system.

Figure 16 - Marzano and Kendall: Levels within the Cognitive System



Source: Marzano and Kendall, (2007, p. 13).

By applying Marzano's and Kendall's theories to examination questions, it can be suggested that students approach examinations and individual questions with differing goals and with varying degrees of motivation and self-belief (the self-system); they also bring differing abilities to design strategies for answering the question (the metacognitive system). Students who can articulate a clear understanding of the steps required to answer a complex examination question show evidence of a well-developed and engaged metacognitive system, and apply relevant knowledge and managing different cognitive processes, whereas those who are confused and disoriented by the question, and are unable to construct effective answering strategies do not exhibit the same levels of metacognitive control.

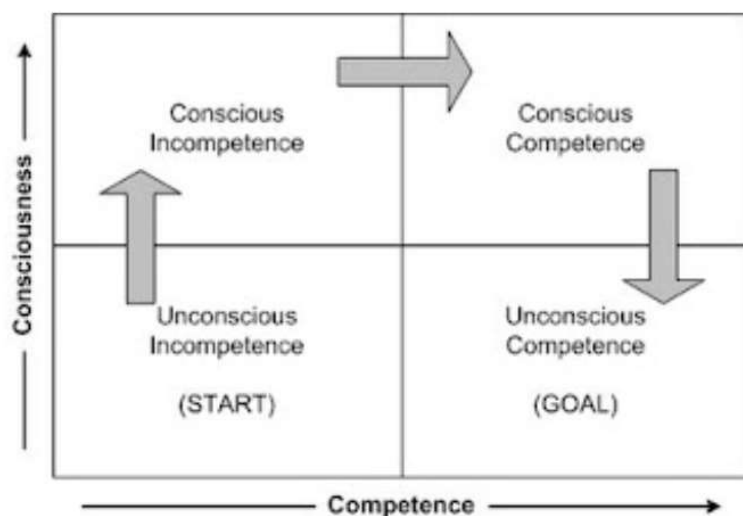
Marzano and Kendall brought significant insights to the classification of learning tasks. They showed that difficulty is not a simple function of the demands of the task alone, but of the interplay within a complex system between the task's demands and the familiarity – that is,

the knowledge and experience – of the individual. They also showed that the working of the cognitive system, which processes knowledge and relevant information, is under the control of the metacognitive system, which is in turn under the control of the self-system. Marzano’s and Kendall’s theoretical insights form the basis for the analysis and discussion of students’ voiced opinions and experiences in Chapter 7.

3.2.3 Noel Burch Competence Model

Marzano’s and Kendall’s New Taxonomy explored the interface between knowledge and familiarity. In a similar way, but approaching the question from a practical and behavioural angle, Noel Burch, working for Gordon Training in the early 1970s, created a competence model that classified four stages of learning any new skill (Figure 16).

Figure 17 - Noel Burch Competence Model



Source: Derived from Reilly (2012), online

Burch’s model (Figure 17) suggested that all learners follow a predictable and sequential route through the acquisition of skills and knowledge. As students start to learn a new skill, they are at first unaware of how little they know, and they tend to underestimate the complexity of the learning task. This is represented as ‘unconscious incompetence,’ in the lower left quadrant.

How long they stay at this stage depends on their awareness and their motivation to learn. As

students become more aware of their own skills or learning deficit, they move to the state of 'conscious incompetence' (upper left quadrant): at this point, they can receive instruction. Further skills and knowledge acquisition can be hard-won: students start to become aware of how to do something (but it requires conscious effort), as they move through a state of 'conscious competence' (upper right quadrant). Eventually, students have had so much practice at performing a skill – or they become so expert in understanding an idea or concept – that it becomes “second nature” and they can perform it without conscious thought, possibly while doing something else at the same time. This is a state of 'unconscious competence,' as shown in the lower right quadrant of the model.

The Burch competence model has gained a good deal of application within education and training contexts, but it has been little discussed academically (ERIC revealed no entries; Google Scholar only 7 citations)⁴¹. This study is, therefore, innovative in applying insights gained from the Burch competence model to an academic context, and through so doing further strengthens the links between professional and academic literature on learning and teaching. Clear links between this model and Marzano's and Kendall's New Taxonomy can be observed: the increasing levels of competence in Burch's model map loosely onto Marzano's and Kendall's knowledge domains and levels of processing within the cognitive system; and the levels of consciousness have connections with Marzano's and Kendall's self-system and metacognitive system.

There are also important differences, however. Burch's model is a handy visualisation of a learning process which is particularly applicable to the learning of psychomotor skills that can become habitual. However, many complex learning tasks in the classroom or the academic sphere are unlikely ever to become fully unconscious, since they rely on the application of (what may have become) a well-practised skill set to a series of new contexts or unfamiliar

⁴¹ Searches made on 21 December 2021, using the terms “competence model”, “Burch”, and “four stages of competence”.

problems. In this respect, even highly able learners are likely to remain at the 'conscious competence' stage, where their metacognitive system is still actively and consciously engaged and managing their cognitive system. To give an example from a mathematics examination question, it is likely that a competent student will be able to recall and apply a well-known formula without much, if any, conscious thought (to calculate the area of a circle or a square, for instance). However, it is likely that the student will then need to engage conscious thought and – engaging the metacognitive system and interacting with the cognitive system – construct a method, if the question then goes on to ask them to apply their understanding to solve an unfamiliar problem (for instance, to determine the relationship between the areas of a circle and a square of the same diameter). Like Bloom's taxonomy, Burch's model gives a framework, a way of describing the observed world of learning. Marzano's New Taxonomy, on the other hand, gives a theory, which allows predictions of behaviour in learners (Marzano and Kendall, 2007).

3.2.4 CRAS Scales of Demands

To classify demands within examination questions, Edwards and Dall'Alba (1981) created a Scale of Cognitive Demands for use in grading science examination questions. They proposed four sub-scales – complexity, openness, implicitness, level of abstraction – each of which had levels. Complexity, for instance, had 6 levels, ranging from '1: simple operations', through '3: understanding, application or low-level analysis' to '6: synthesis or evaluation' (Edwards and Dall'Alba, 1981, p. 159). These levels appear similar to those in Bloom's (1956) Taxonomy but, like Marzano and Kendall (2007), they present an analytical model of demands rather than a hierarchical taxonomy. Pollitt *et al.*, undertook research for the Qualification and Curriculum Authority (QCA),⁴² which looked to define and explain the concept of demand 'for the

⁴² QCA was the UK Government's regulator for examinations. It was formally dissolved by the UK Government in 2010, when its regulatory powers were transferred to the Office of Qualifications and

description and evaluation of examination standards' (2007, p. 166). Pollitt *et al.*, (2007)

modified Edwards' and Dall'Alba's (1981) scale in order to make it applicable to subjects other than the sciences. The scales seek to measure complexity, resources, abstractness, and strategies, hence "CRAS" scales (Table 1).

Table 1 - CRAS Scales of Demands

	1	2	3	4	5
Complexity The number of components or operations or ideas and the links between them.		Mostly single ideas or simple steps. Little comprehension, except that required for natural language. Few links between operations.		Synthesis or evaluation is required. Need for technical comprehension. Make links between cognitive operations.	
Resources The use of data and information.		More or less all and only the data / information needed are given.		Student must generate or select the necessary data / information.	
Abstractness The extent to which the student deals with ideas rather than concrete objects of phenomena.		Mostly deals with concrete objects.		Mostly abstract.	
Task strategy The extent to which the student devises (or selects) and maintains a strategy for tackling the question.		Strategy is given. Little or no need to monitor strategy. Little selection of information required.		Students need to devise their own strategy. Students must monitor the application of the strategy.	
Response strategy The extent to which students have to organise their own response.		Organisation of response hardly required.		Must select answer content from a large pool of possibilities. Must organise how to communicate response.	

Source: Pollitt *et al.*, (2007, p. 186).

Examinations Regulation (Ofqual). Both QCA and Ofqual have been colloquially termed the 'examinations watchdog'.

Pollitt *et al.*, were very clear about the links between the input of the examination system (the question paper) and the outcome (the grade a student achieves): 'it is generally assumed that the score depends on two factors – the ability of the student and the difficulty of the questions' (2007, p. 193), but it can only be ensured that more able students gain higher grades if all students respond predictably to the demands of the questions posed. They noted other studies (for example, Ahmed and Pollitt, 1999; 2007; Crisp and Sweiry, 2005, all cited in Pollitt *et al.*, 2007), which showed that differences in the performance of individual students on individual questions 'depend at least as much on the presence or absence of various features in the stimulus question, as on the amount of difficulty the examiners intended' (Pollitt *et al.*, 2007, p. 193). Because these distracting factors, presentational and contextual, can distort the straightforward translation of question demand into difficulty, as experienced by the student, they are a threat to the validity of inferences made from examination results. In their discussion about how aspects of demand are the direct cause of difficulty for students answering questions, Pollitt *et al.* (2007) revised the earlier work of Pollitt, *et al.* (1985), in which three sources of difficulty were identified: subject/concept difficulty, process difficulty, and question (stimulus) difficulty:

'We now consider difficulty resulting from the concepts in a subject as aspects of demand; in the CRAS scheme they are rated under 'abstractness' or 'complexity'. Similarly, difficulty arising from the psychological processes the students are asked to carry out is rated as demand in the scales for 'strategy', 'resource', and 'complexity'. For these categories it is fairly simple: more demand quite directly causes more difficulty, and this can be observed as lower scores for students' (Pollitt *et al.*, 2007, p. 193).

Strategy was divided into two different parts, 'since exams might differ in the balance of the demands they make on devising strategies for solving problems and on planning how to communicate the answer once it has been found' (Pollitt *et al.*, 2007, p. 186). Since 2007, the CRAS scales have been used in studies that have compared the demands and grade standards in subjects as diverse as economics (Greatorex *et al.*, 2013); life sciences (Dempster and Kirby, 2018); vocational qualifications (Novakovic and Greatorex, 2011); and mathematics (Tan *et al.*,

2017). Johnson and Mehta have further observed that the CRAS framework is 'essentially qualitative in nature and can be used to profile the nature of cognitive demands for individual users' and they cautioned that it is 'not possible to combine ratings to reach an overall level of demand' (2011, p. 31).

Many existent studies of demand and difficulty have been authored by researchers from the professional assessment community; others are by university academics. Examiners were engaged in the testing and refinement of the CRAS scales. However, the involvement of students and teachers – that is, stakeholders from outside the assessment community – in rating the demands of examination questions has not previously happened. 'Yet,' as Pollitt *et al.*, observed,

'There are good arguments for using groups other than examiners. Teachers, who prepare students for the examination and are not practised in the arts of question writing, may be in a better position to judge how students will be challenged by a particular feature than examiners who recognise it from past papers. Of course the students themselves are even more likely to understand how demands really operate' (Pollitt *et al.*, 2007, p. 188).

The almost throwaway line ('of course the students themselves...') implies almost a shrug from the authors: as employees of an examination board, they possibly realised how unlikely it was that students would be asked to contribute their understanding. However, Pollitt *et al.*, (2007) here made the case for a wider range of participants in the evaluation of the demands of examination questions. This is a key objective of this study.

In this study, the observations of students will be related to the Marzano and Kendall New Taxonomy (2007), because it provides a more multi-dimensional model of cognition, and students' explanations will be examined in the light of the conceptual model already outlined, derived from cognitive load theory. The relation of student responses to established theory is an innovative feature of the current study.

Turning to the relationship between the 'demands' of an examination question and its resultant difficulty for the student, the general assumption was presented, based on research

literature, that the student's score in an examination depends on 'the ability of the student and the difficulty of the questions' (Pollitt *et al.*, 2007, p. 193). It has already been noted, however, that presentational and contextual factors in the question may interfere with this direct relationship, in ways that may not be predicted or intended by examiners; these factors will be investigated in the empirical studies. The student's self-system (Marzano and Kendall) and self-efficacy (Bandura, 1997) may also play a part in controlling their ability to apply their knowledge and understanding to meet the demands of the examination question. It is possible to deduce that the difficulty of a question, for each individual student, therefore, may be a function of the demands of the question and a combination of the expertise and self-efficacy of the student.

3.3 Validity considerations underpinning high-stakes assessments

There is a wealth of literature on validity and validation. A search for the term ‘validity’ in the Education Resources Information Center (ERIC)⁴³ yielded 48,860 results, of which 32,835 are peer-reviewed, 6,318 of them in the five years between 2017 and 2021. The same search with Google Scholar⁴⁴ yielded ‘about 3,680,000’ results, of which ‘about 913,000’ are since 2017.

There are differences and potential tensions between the ways in which validity is understood and operationalised within quantitative and qualitative research, as well as between academic and lay meanings of terms. It is important to be clear what is being discussed. Kane pointed out that, ‘how we choose to use a term depends on what we want to do with it.’ He continued, offering a personal contextualisation:

‘I think of validity as the extent to which the proposed interpretations and uses of test scores are justified. The justification requires conceptual analysis of the coherence and completeness of the claims and empirical analyses of the inferences and assumptions inherent in the claims’ (Kane, 2016, p. 198).

This definition is different from the more general meaning in common use (for example, in the Merriam Webster dictionary⁴⁵, where validity is defined as ‘the quality of being well-grounded, sound or correct’). Or, in terms of statistics, a loose definition of what some authors have called “face validity” (for example, McCormick and James, 1983; Fautley and Savage, 2008) is often derived from Kelley, in that the validity of a measurement tool is ‘the degree to which the tool measures what it claims to measure’ (1924, p. 194).⁴⁶

Shaw and Crisp reviewed the history of thinking on validity: early authors considered validity to be ‘a static property captured by a single statistic, usually an index of the correlation of test scores with some [other] criterion’ (Shaw and Crisp 2011, p. 11). This approach is seductive in its simplicity: validity is understood as a measure of error, that is, the distance between the

⁴³ ERIC – eric.ed.gov – an online digital library of education research and information sponsored by the Institute of Education Science (IES) of the United States Department of Education, accessed 20.09.2021.

⁴⁴ <https://scholar.google.com>, accessed 20.09.2021.

⁴⁵ www.merriam-webster.com accessed 24.09.2021

⁴⁶ See, for example, en.m.wikipedia.org accessed 24.09.2021

measured value of the test and the “true” value. An evident problem with this content-based approach, however, is that there is no way to calculate the “true” value. Cureton, in the 1950s, took a criterion-referenced view, stating that ‘the essential question of test validity is how well a test does the job it is employed to do’ (1951, p. 621). This pragmatic definition is attractive, and it may work for small test situations; in the context of a large-scale high-stakes public examination, however, it is not at all clear what ‘the job’ is that such an examination ‘is employed to do.’ There are too many different purposes served by an examination grade for this definition to work well. Cureton’s view, though, shows in part the shift from thinking about validity in terms of a property of a particular test to a judgement about how well the test fulfils the functions expected of it. The “face validity” idea, that validity is about whether a test ‘does exactly what it says on the tin⁴⁷,’ is still surprisingly persistent.

There has, however, been a shift from talking about ‘validity’ in a fixed way, related to a view of scientific realism, to discussing a process of ‘validation.’ In published writing about educational and psychological assessment, criterion-based validity thinking gave way to ‘construct validity,’ a term coined by Meehl and Challman in 1954 and developed by Cronbach and Meehl (1955). A ‘construct’ is an attribute of those taking the test (“intelligence,” for example, or the ability to solve quadratic equations), and construct validity aims to evaluate the extent to which a test is an adequate measure of that construct. Unlike criterion validity and content validity, construct validity focuses on the purpose of the assessment. In order for construct validity to have meaning, test developers must have a clear understanding of the construct they wish to assess. Cronbach (1971) compared research into validity to the evaluation of a scientific hypothesis. Kane (2016) took this scientific analogy further, and proposed that an interpretative argument should be advanced as well as a validity argument. This interpretative argument concerns the inferences and assumptions that will follow from

⁴⁷ Television advert for Ronseal woodstain, 1994: <https://www.creativereview.co.uk/does-exactly-what-it-says-on-the-tin/> accessed 10.07.2023

the test scores to the uses to which these scores are put. Having accepted the consensus view of construct validity, this study follows Kane's argument-based approach.

In terms of GCSE examinations, a grade derived from an examination – whether it be an “A*” grade or a grade “9” – is an abstract label; it means nothing on its own. These letters and numbers must stand for something, and they need to convey meaning to the user: they require context and interpretation. For example, in order to evaluate what a Grade 9 result in GCSE Mathematics might mean, it would be helpful to start with an understanding of the content of the GCSE Mathematics specification, to know that the grading system ranges from 1 to 9, where 9 is the highest grade, and that Grade 9 was awarded to only 3.7% of the 720,098 students who took the examination in 2019.⁴⁸ The notion of validity provides an evaluation of the interpretative claims that may be made for a grade or result.

Validity, then, is an abstract concept rather than an objectively measurable property (such as height or mass). Newton gave the standard or “consensus”⁴⁹ definition of validity within the academic and research community:

‘The 1999 Standards for Educational and Psychological Testing defines validity as the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests’ (2012, p. 1).

Camargo *et al.*, (2018) confirmed this consensus and showed, through their Delphi study, that while there is still room for disagreement among experts about several aspects of validity (the outcome or judgment) and validation (that is, the process of establishing validity), there is general agreement around the definition that, ‘Validity is the degree to which collected evidence, theory and argumentation support inferences based on observed scores’ (Camargo *et al.*, 2018, p. 149). Here the addition of the phrase ‘and argumentation’ is important. Kane (2006) perceived validity not as a quantifiable outcome, but in terms of a validation process

⁴⁸ <https://schoolsweek.co.uk/gcse-results-2019-mathematics/> accessed 29.09.2021

⁴⁹ Newton (2012) gave this as the consensus definition of the educational and psychological measurement and assessment (EPMA) communities in North America, and stated that it also has substantial international currency.

that assembles an extensive argument – that is, a justification – for the claims that are made about an assessment. Kane also explained that:

‘An interpretation is said to be ‘valid’ if it is supported by appropriate evidence (Cronbach, 1971; Kane, 2006; Messick, 1989; Mislevy, Steinberg, and Almond, 2003). The interpretation is not ‘valid’ if the proposed interpretation is not justified’ (2016, p. 198).

The concept of validity, then, is bound up with, and inseparable from, the interpretations to which the results of a test are put: it is not the test itself that can be described as valid or invalid, but the interpretation of the test or examination result in the overall context of the test or examination. This is a vital refinement, and it has implications for the ways in which validity is discussed.

As Newton pointed out, it is important to set out some ground rules for talking about validity within the context of educational assessment and research:

‘First, it is bad practice to talk about validity as though it were a property of a test. Second, it is good practice to describe validity as though it were a property of an interpretation. Third, it is good practice to describe validity as a unitary concept. Fourth, it is good practice to define construct validity as the unifying essence of all validity’ (2012, p. 2)

Newton’s succinct summary, arriving at ‘construct validity as the unifying essence of all validity,’ consolidated the position arrived at by Messick, where he had defined validity as:

‘An integrated evaluative judgement of the degree to which empirical [*sic.*] evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment’ (1989, p. 13, emphasis in the original).

Recent studies, such as Hopfenbeck, (2020); Addey, *et al.* (2020); and Zhou, *et al.* (2020), have sought to re-evaluate this consensus view of assessment in different ways. In particular, although the study by Addey *et al.*, (2020) sub-titles itself as ‘rethinking Kane’s argument-based approach,’ they actually sought to clarify “whose” context is being considered when validity is discussed, as assessments become increasingly large-scale and international in nature. When assessments cross national and cultural borders, their use, context, and

interpretation can change dramatically. The global COVID-19 pandemic – which formed the backdrop to the main empirical study of the present research – also led to changes in the administration of high-stakes assessments and, through so doing, contributed to further reflections on validity and validation processes (see amongst others, Hopfenbeck, 2020; and Zhou *et al.*, 2020).

3.3.1 Validity in relation to GCSE examinations

Newton, having clarified the consensus definition of construct validity, remarked powerfully that ‘validity is a property of an assessment-based decision-making procedure’ (2012, p. 18). In the case of GCSE examinations, this presents some issues. The purposes to which a student’s grade may be put in the future are many and varied. Purposes may include:

- For the student, progression to the next stage of education, or selection for employment in the future;
- For the teacher, when amalgamated with the results of the rest of the class, a professional discussion around methods of teaching and learning, training needs or an appraisal judgement;
- For the school, contribution towards a whole-school judgement of progress, an assumption of school effectiveness by prospective parents, or an inspection grade.

Not all examiners appear to have studied validity theory, however, and misconceptions abound in formal and informal writing about validity in relation to examinations. Examples of “loose talk” abound; here are three:

AQA, 2019: ‘validity is about whether or not, and there's lots of definitions of validity here, but for us it's about whether or not you're assessing the right thing in the right way.’ ... ‘Now, in an ideal world, for the most valid form of assessment, you would

want to cover all the content every year.’ (Dave Mellor, Director of Assessment and Curriculum at AQA, interviewed by Craig Barton, mathematics teacher and author).⁵⁰

Oxford Assessment:⁵¹ ‘A valid exam is one that measures student performance accurately in the subject being tested.’

Ofqual, 2018:⁵² ‘The most significant benefit [of teacher involvement in developing examination papers] was perceived to be the ability of teachers to use their applied assessment skills alongside their subject and student knowledge to make exams as valid and reliable as they could be.’

All these statements – from a professional examiner, an examination board, and the examinations regulator – write of a ‘valid exam,’ or otherwise imply that validity is a property of the test and not the inferences made from the results of the test. Mellor, in atomising ideas about validity and stating that, ideally, ‘you would want to cover all the content every year,’ also strays into ideas about content validity. To pick up on this is not merely arguing about semantics. The idea that validity is simply a matter of whether an assessment tests what it sets out to test, has been regarded as inadequate for more than 70 years. It is an unfortunate feature of the development of validity theory, and its resultant literature, that it has left a litter of misconceptions and broken definitions in its wake.

Returning to the consensus principle that validity is about the interpretation of the results of high-stakes assessments, Ahmed and Pollitt explained that the validity interpretation of an examination result rests on the straightforward principle that,

‘The test constructors’ task is to ensure that the questions and mark schemes that they write deliver scores for students that show as accurately as possible how much and how well they have learned. To the extent that the assessment fails to do this, the potential for interpreting the results validly will be threatened’ (2011, p. 260).

⁵⁰ <https://www.aqa.org.uk/inside-exams-podcasts/series-2-episode-3> accessed 22.09.2021

⁵¹ <https://oxfordaqaexams.org.uk/why-oxford-aqa/fair-assessment/achieving-a-valid-exam> accessed 22.09.2021

⁵²

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/690002/Teacher_involvement_in_developing_exam_papers_Findings_from_our_call_for_evidence_.pdf accessed 22.09.2021

Ahmed and Pollitt set out a view that the concept of a test has its own inherent validity – that is, every examiner sets out to create a test that accurately measures the expertise of a student in such a way that the results can be validly interpreted – but that threats to this validity creep in as the examination process develops through the unintended consequences of pragmatic steps:

‘At the start of an assessment process, when the eventual test is no more than an idea in the constructors’ minds, the potential for good quality assessment is very high, but at each step compromises and poor choices may reduce this potential’ (2011, p. 260).

Ahmed and Pollitt defined three threats to validity. The first of these is that ‘the questions may not elicit from students the kinds of behaviour that the examiners want to evaluate’ (2011, p. 260). Their other two threats to validity are concerned with the reliability of marking, which though important, are beyond the scope of this study.

Ahmed and Pollitt (2007) had earlier studied the effect of manipulating science questions on how well students were able to answer the questions. Their concern was that presentational aspects of examination questions could deflect the attention of students, and thereby represent a threat to the fidelity of the assessment, and hence to validity. They summarised their argument, stating that,

‘Any aspects of a question that get in the way of the students ‘doing the things we want them to show us they can do’ are a potential threat to fidelity, and so to validity. Setting questions in real world contexts is one such potential threat’ (Ahmed and Pollitt, 2007, p. 202).

The reference to “fidelity” here is to a freshly-coined term “construct fidelity”, which the authors used to describe the contribution of questions to overall validity. For them, the variability introduced by these contextual factors – which may not only affect different students in different ways, but also interfere unpredictably in the question-answering process – is unwelcome.

Setting questions in real world contexts, however, is a frequently encountered feature of examinations, and it is explicitly included as part of the examination specifications, even in a

more abstract subject such as mathematics (see, for example, OCR, 2021, where real-world illustrations are given in the specification). Little and Jones (2010) cited some possible reasons for this practice. On the one hand, mental scaffolding might provide students with some assistance; on the other, real-world contexts might appear to complicate the task, because they may assume knowledge of the context as well as of the mathematics, which may add to the cognitive load of the problem.

Spalding (2011, p. 12) also considered ways in which ‘presentational and structural considerations affect candidates’ performance,’ and concluded with some comments on the validity aspects of examination paper construction. Her central point was that, if elements intrude unpredictably between the intentions of the examiner and the performance of a candidate, the validity of inferences that can be made from the test result become significantly weaker:

‘If candidates misread questions, misinterpret the requirements of the given task, or miss questions which they are capable of answering due to either presentation or placement in the paper, then the paper is no longer accurately assessing ability... If the paper is no longer solely assessing what it is intended to assess, then it loses validity. An examination paper is not simply as good as the questions within it; presentation and structure make it greater than the sum of its parts’ (Spalding, 2011, p. 13, emphasis added).

Leaving aside for one moment Spalding’s apparent slip (writing that ‘the paper... loses validity’), the points she raises about possible sources of distraction and bias caused by the presentation of the questions are vital; they are subsequently considered in Chapters 5 to 7 of this study.

Despite the clear evidence that that considerable time and resources have been expended by examination boards on research into crucial aspects of examinations, including validity and question design, there is a near total absence of literature on examiners’ engagement with research on any aspect of examinations. A website search for this study revealed that none of the three major examination boards in England (OCR, AQA, and Edexcel) mentions research in

their online examiner recruitment information, whilst details of training extended only as far as marking papers and standardisation. These are undoubtedly important areas for examiners, but the lack of any information on engagement with research, questions of validity, or the technical and evaluative side of question setting and assessment is startling.

In 2008, the Qualifications and Curriculum Authority (QCA) undertook a review of question paper setting and senior examiner training. Although training was observed and reviewed, references to research and validity are conspicuously absent from the report, and in the 15 years that have elapsed since 2008 no further reviews have been undertaken. Moreover, it is also the case that, in the years immediately preceding 2008, there was clearly no expectation for senior examiners to have any familiarity even with the output of their own awarding bodies' research arms (Spalding's paper was published by the Centre for Education Research and Policy, AQA's then research arm; Ahmed and Pollitt both worked for Cambridge Assessment, the research division of OCR). QCA's report suggests a light touch approach to training: most of it appeared to have been conducted by experienced examiners passing down their wisdom, or through examiners finding out for themselves. With regard to the important aspect of ensuring that examination papers have a wide enough range of demand in their questions – so as to ensure that individual papers discriminate sufficiently between less and more able students – the report noted, for example, that,

'Some guidance was provided on ensuring that question papers have an appropriate range of demand, but this tends to be developed through examining experience, rather than through training meetings and materials.... [and that] issues affecting demand often were not addressed through awarding body training meetings or materials, and usually awareness of them is developed through acting in the senior examining role and the experience of leading an examination' (QCA, 2008, pp. 3, 12).

No evidence was discovered that senior examiner training has yet systematically taken account of research that would help examiners to improve the quality or effectiveness of their assessments, or which would improve their understanding of the validation processes that should be undertaken to ensure the stability, quality, and reputation of high-stakes

examinations. More active engagement between the educational assessment research community and professional assessors, including senior examiners, might help the concept of validity in high-stakes assessments to become a practical activity and concern, linked to the improvement of the quality of assessments, rather than a largely theoretical concept.

3.4 Student voice

There are two complementary rationales for including the voices of students in a discussion upon assessment. These views come from “outside” and “inside” the educational assessment community. The outside or external view, is defined in Article 12 of the United Nations Charter on the Rights of the Child which states that:

‘Parties shall assure to the child who is capable of forming his or her own views the right to express those views freely in all matters affecting the child, the views of the child being given due weight in accordance with the age and maturity of the child’ (United Nations, 1989, online).

This view has been picked up in academic literature:

‘The results of national assessments such as GCSEs have a major impact on their future life trajectories, and so students’ views on this topic should be considered’ (Barrance, 2019, p. 567).

‘Student voice is a normative project and has its basis in an ethical and moral practice which aims to give students the right of democratic participation in school processes’ (Taylor and Robinson, 2009, p. 161).

The inside or internal view (see, for example, Lundy, 2007; Barrance, 2019) advances that taking account of students’ voices can improve aspects of the validity of assessments. That is to say, if students are inside the system that is designing the assessments they will take then, by giving feedback on the format and content of test items as they are created, it is possible that the resulting examinations may more reliably and accurately test their knowledge, understanding, and skills. Where the views of students come from outside the system of test developers, the test is created “blind”, with little real idea of its ability to perform its valid

functions. As discussed in Section 3.3, this should be a matter of concern particularly where questions and examinations are not pre-tested and do not come from item banks, as in the UK.

Some studies have acknowledged that standard approaches to assessment may be missing something important:

‘Governments worldwide have introduced assessment reforms, but few countries have included research as part of the process. Even fewer have examined students’ perceptions of such reforms’ (Hopfenbeck, 2019, p. 255).

‘It is timely to challenge persistent and well-critiqued approaches to voice that maintain students in power relations where they report on schooling practices but exert limited influence on school decision-making or classroom relations’ (Charteris and Smardon, 2019a, p. 106).

‘The field of educational assessment has built strong technical foundations, but we can be myopic when the big education questions are asked’ (Gray and Baird, 2020, p. 137).

Periodically, the promise of a new beginning is heralded with regard to engagement with student voice. So, for example, Mitra (2001, p. 105) wrote of a ‘new awareness of the reciprocal nature of learning’ as teachers and students listen to one another. As Bourke and Loveridge (2018) observed, however, such optimism rarely leaves a lasting mark on educational or research practices. Barrance (2019, p. 566) noted that ‘the exclusion of students’ perspectives [from the debate about assessment reform] is particularly problematic if we consider... that students are not passive recipients of policy’ but are intimately bound up in the enactment of such reforms. This view amplifies similar concerns explored by Ball *et al.*, (2012); Barrance and Elwood (2018); and Elwood (2012). Sandoval and Messiou (2020) reviewed 28 studies published between 2004 and 2018 in which students were involved as researchers; these studies were carried out across four continents. The focus (and number) of these studies was: learning (7 studies); engagement (6 studies); bullying and behaviour management, school experiences (3 studies each); homework, transition, technology, school culture (2 studies each); and school meals (1 study). Not one of them studied examinations. This present study therefore speaks into this gap in research literature. Braun and Clarke (2022, p. 120), counsel researchers using reflexive thematic analysis within a qualitative

paradigm to be wary of the urge to ‘establish the gap,’ since this ‘reproduces a positivist-empiricist idea of research as truth-questing,’ where any gaps in knowledge *need* to be filled. Against this, they acknowledge that the fact that there *is* a gap provides an opportunity for exploring the lived experience of students through a qualitative study in a way that, in line with a more qualitative paradigm, makes a contribution to the rich tapestry of knowledge and understanding. The main rationale for this present study, therefore, is not to establish and fill a known gap, but to make an argument for the inclusion of student voice in the evaluation of high-stakes assessments.

Despite this dearth of researched engagement, the benefits of asking students about their experience of difficulty in examinations appear self-evident – they have a right to be consulted, since these matters affect them deeply, and they may be able to help improve the validity of the very assessments that they are about to undertake. Questions are then generated about how and in what contexts the student voice is sought, and whether – having been sought and heard – it is in fact listened to. In other words, why are students asked; what are they asked to speak about; and what difference does their voice make? The reasons that the voices of students have not been sought in studying the design and effects of examinations may be varied. Three main reasons have been advanced as to why student voices have not been sought.

First, the concept and use of student voice has grown up within a neoliberal, modernising educational context that is, in general, far removed from the more traditional ideological stomping ground of test developers (see Pearce and Wood, 2019; subsequently discussed herein). This modernising educational agenda is often politically-driven and detailed; political directives towards examination boards, on the other hand, have tended to operate at the level of broad policy changes rather than detailed guidance.⁵³ In the United Kingdom, some of the

⁵³ For example, the UK Government mandated the change of GCSE grading systems from A*-G to 9-1; or the suspension of public examinations during the summer seasons of 2020 and 2021 during the COVID-19 pandemic. Examination boards were left to determine the details and workings of these policies.

incentive for the use of student voice in education has been mandated by politicians, starting with Blair and the New Labour government in 1997⁵⁴ and continuing through changes of government and political leadership through the next two decades. The Department for Education (DfE, 2014a) provided statutory guidance to schools on listening to and involving young people, stating that schools must identify ‘how best to provide opportunities for pupils to be consulted on matters affecting them or contribute to decision-making in the school’ (p. 1). The Department for Education further observed that the benefits to students of making active contributions to decision making include ‘increased confidence, self-respect, competence and an improved sense of responsibility’ as well as ‘increased motivation and engagement with learning’ (p. 2). However, whereas governments have directed and encouraged schools to engage with student voice, no such encouragement has been directed towards examination boards, which may explain why examination boards have rarely felt the need.

Secondly, and linking to the first reason, students have traditionally been regarded as the subjects of tests and examinations, and paradoxically have not been truly regarded as having a stake in the assessments they are undertaking. To parody Abraham Lincoln’s famous phrase in the Gettysburg address (1863), these tests are *of* the students, not *for* the students, and certainly not *by* the students.

Taking the principle of student voice to its logical extreme, where students are involved in every aspect of research, is the movement called Youth Participatory Action Research (YPAR): see, for example, Cammarota and Fine, (2008); Rodriguez and Brown, (2009); and Zaal and Ayala, (2013). Through the involvement of young people in all aspects of research, the YPAR

⁵⁴ Tony Blair delivered his ‘Education, Education, Education’ speech to the Labour Party Annual Conference in Brighton, 1997, at the start of the New Labour government that saw a number of seminal policies prioritising the needs of children and young people. These included *Learning to Listen: Core Principles for Involvement of Children and Young People* (DfES, 2001) and *Every Child Matters* (DfES, 2003). These initiatives came partly in response to the United Nations Charter on the Rights of the Child (United Nations, 1989), as was explicitly acknowledged in the DfE (2014) guidance *Listening to and involving children and young people*.

movement 'takes seriously youth contributions to tackling issues that affect their lives and communities, rather than viewing youth as problems in need of intervention' (Zaal and Ayala, 2013, p. 161). Young people identify the topics to be researched; they design the research instruments to be used; they carry out the research and analysis; and they present their findings. This movement is strongly linked with social justice concerns in Latin America and the United States, including improving educational opportunities within indigenous and black communities. High-stakes examinations have not so far been a focus for YPAR studies. The YPAR approach is totally immersive in its student-led focus. This study takes note of the strengths of YPAR in acknowledging the feelings and interests of students, but it takes a more traditional researcher-led approach.

Thirdly, students speak with diverse voices and from a range of viewpoints – it is possible to refer in the singular to the student voice, but it would be better to speak of student voices in the plural. This divergence and variety of opinions can make students' views hard to assimilate, as Dockerill (2018) notes in his chapter on "forgotten voices". Some of the students' views may also be uncomfortable to education professionals because they may challenge established hegemony and practice. Burton (1995) demonstrated that secondary school students, when consulted, revealed experiences and understanding that were in direct contradiction to positions presumed by the adults who had constructed standard curricula and authored published teaching resources. The title of Bragg's (2001) article 'Taking a Joke: Learning from the Voices We Don't Want to Hear' summed up this discomfort. More recently, Bourke and Loveridge (2018) argued that this challenging function is precisely the point of engaging with student voice, but they acknowledged that this approach runs counter to the prevailing trend.

The questions already debated have sprung from a consideration of the different functions and approaches of student voice. Hall (2017) and Charteris and Sardon (2019b) presented dichotomies of approaches. Hall (2017, p. 181) distinguished between what she termed 'everything' and 'nothing': on the one hand (drawing on Rudduck and Fielding, 2006) the

'transformational' and, on the other hand, the 'tokenistic.' What Hall and also Rudduck and Fielding meant by this dichotomy is that, at the 'nothing' or 'tokenistic' extreme, students are consulted merely as a formality, on matters that have already largely been decided by others (the adults). Students' views here may be interesting, but they are unlikely to cause the format or content of their education to change significantly. At the other, radical, extreme, 'everything' is up for discussion, so that there is an opportunity for students' views to be 'transformational', in that they are likely to lead to real change and may even contribute to the design of particular patterns of learning or assessment.

It is easy to see how these different approaches may have arisen. As stakeholders in their own education, the voices of students arguably have a place in any evaluation of the effectiveness of this education; it would be hard to imagine any commercial venture surviving long if it did not wish to ascertain and take account of the voice of its consumer base. But there is a longstanding debate within education, on how much power to give to the voice of the consumer: whether they are being asked to validate the professionals' own views (a rather safe position), or whether the intention is to seek to draw them into transformational dialogue about their experience – their learning – which feels much more risky. Within student voice literature, Charteris and Smardon characterised these two different positions as 'democratic contribution or information for reform' (2019b, p. 93).

Kelly provocatively captured the question deep at the heart of this dichotomy: 'Who cares what the kids think?' (2019, p. 1). Referencing Lodge (2005), she identified four types of engagement with student voice: quality control; students as a source of information; compliance and control; and dialogue. Lodge's/Kelly's taxonomy in effect places these four positions on a conceptual continuum between tokenistic consultation and transformative dialogue. The importance of their classification is to illustrate that the majority of studies using student voice belong in the first two categories – where students are used for quality control or as sources of information; in contrast, the aim of this study is to bring students' voices into

an academic and educational forum where they may effect real change: to offer opportunities for dialogue. In a similar manner to Kelly's, Mitra, building on her own previous work from 2003-9, gave another classification, and useful definitions for each class of student involvement (Figure 18):

Figure 18 - Three-fold Classification of Student Voice



Source: Mitra (2018, p. 474).

Mitra's definitions were that, at the level of listening,

'Adults seek student perspectives and then interpret the meaning of the student data,' [whereas collaboration is when] 'adults and youth work together. The adult tends to initiate the relationship and ultimately bear responsibility and the final say on group activities and discussion.' [In positions of leadership,] 'students assume most of the decision making authority and adults provide assistance. Most examples exist outside of the auspices of the school' (Mitra, 2018, p. 474).

Mitra noted, citing Fielding (2001) and Lundy (2007), that while the UK has probably more examples of "youth participation" (the United Nations Convention on the Rights of the Child term for student voice) than anywhere else, 'often such participation is merely tokenistic or symbolic rather than manifesting as a true act of collaboration with young people,' and she styled the attitude of many of these approaches as 'begrudging' (Mitra, 2018, p. 475).

The students that are consulted within existent studies tend to be in either primary school or in higher education⁵⁵ and an evaluation of these studies appears in section 3.4.2. There are fewer published examples of the secondary school student voice in relation to assessment. Motives behind researchers seeking students' involvement are complex, but possibly not often student-centred. Educators, as shown below, tend to want to hear an affirmative rather than a contrary perspective when students use their voice. As Mitra *et al.*, argued, 'when developing student voice initiatives, one of the greatest struggles is the role of the adult in these interactions' (2012, p. 104). They meant that the adult can influence both the views expressed and the ways in which they are interpreted and used following the interaction. Mitra warned that 'the promise of voice without actually being heard can lead to increased alienation and disconnection from schooling' (2018, p. 475). It is useful to examine some of the context of this debate, and to explore the difficulties in more depth, because these points bear directly on this study's empirical work.

Pearce and Wood argue that the term 'student voice' has 'taken on many meanings and is used to fill a range of largely ideological purposes' (2019, pp. 113-4). In a wide-ranging evaluative review of published work on student voice, they offer several important observations that may also serve as warnings to would-be researchers of student voice. Pearce and Wood note first that the concept of student voice has originated within contexts that have particular historical and social characteristics. Dominant among these are groups pursuing school reform, as part of 'distinct but aligned neoliberal, neoconservative and managerial middle class groups' (Pearce and Wood, 2019, p. 114), or what Apple, *et al.* term the 'conservative modernisation' of education (2009, p. 10). This urge towards modernisation seeks to harness student voice as part of a discourse that aims to raise educational outcomes in standardised assessments. This aim has been correlated with economic growth (Hanushek and Wößmann, 2008; and Tikly, 2011). Teachers and schools in such regimes can come under

⁵⁵ As revealed by searches using search engines including Google Scholar and ERIC, accessed 10 December 2021

‘intense and direct pressure to improve students’ results’ (Pearce and Wood, 2019, p. 115), and in these circumstances the importance given to student voice can materially alter the relationship between teacher and student (Ball, 2003). The democratisation of the student voice, and its acquisition of real power and significance, therefore, carries revolutionary overtones in both social and educational terms.

There are clear implications for education professionals who wish to hear the student voice in order to improve their educational experiences, particularly in the field of educational assessment. If the inclusion of student voice processes is imposed from a managerial or leadership level then, even when it has the clear aim of improving the quality of teaching, the approach may well provoke unease or suspicion from class teachers and unions. Here, for example, is published guidance from one of the largest teaching unions in the UK:

‘Principle 2 – Student voice activities must not undermine teachers’ professional authority and must not compromise other fundamental rights of children and young people...’

‘Any student voice practice that is used to make judgements about a teacher’s professionalism and so has the potential to undermine teachers’ professional authority is unacceptable. Unfortunately, the NASUWT has received examples of schools using student voice to question teachers’ capabilities. Not only is this unacceptable employment practice, it is likely to create suspicion and resistance and undermine any benefits of student voice’ (NASUWT, 2020, online, emphasis added).

There are clear protocols around leaders’ and managers’ observation of lessons within schools in many jurisdictions, but the gathering of evidence using student voice is viewed by some as a way of circumventing these safeguards, particularly since the use of student voice has been aligned with specific school reform agendas. As Mitra noted,

‘A big reason for the reluctance of adults to increase student voice is that the institutionalised roles of teachers and students in school contradict much of what an adult-youth partnership is about’ (2018, p. 479).

To avoid misunderstanding, therefore, and also maximise the positive effects of feedback from students, it is advisable to involve teachers and subject leaders in consultation and discussion

about the intended use of student voice. Mitra's suggested solutions to this deep-seated problem sound radical:

'Cultivating trust and respect, celebrating successes, teaching how schools work, creating a flat power dynamic, building an inclusive community and signalling partnership through visual cues' (2018, p. 480).

Mitra gave examples from youth organisations in Vermont, Boston, and Michigan (USA) where these approaches have been used successfully; however, none of them was in a school.

Applying these principles to a school context in the UK is beyond the scope of this study, but it would surely be possible to embody trust and respect in dealings with students and create a sense of real partnership. In this study, discussions with teachers and senior leaders regarding the role of student voice took place at the design stage of the study; these professionals expressed their eagerness to learn from any of the study's findings.

Extending their historical perspective, Pearce and Wood (2019) argued that different registers of student voice are generated at different times, and that they tend to echo aspects of a prevailing society or hegemony, 'leading student voice initiatives to simply reproduce relationships of power and domination instead of providing alternatives or challenges to the established social order' (Pearce and Wood, 2019, p. 116). Some education professionals, then, are happy to hear the student voice, so long as it tells them what they already think they know. This can feel uncomfortably like an extension of the Victorian notion that children should be "seen and not heard". Kelly classified this aspect of student voice as one of 'compliance and control', in which some account is taken of the rights of students to be included in decisions, and yet the student voice is 'utilized to serve institutional ends' (Kelly, 2019, p. 27). This leads to a certain sceptical pessimism, evident in some writing about student voice, wherein the authors appear almost resigned to see student voice as an amplification or echo of the dominant educational and political discourse (see, for example, Bernstein, 2003, and Pearce and Wood, 2019). This approach devalues the contribution that student voice can make. Other writers, however, appear to urge their readers to listen more intently to the

student voice, seeking to discern ‘the speaking personality... [with] its own timbre and overtones’ (Bakhtin, 1981, p. 434).

The elusiveness of the real voice of students has been a concern for a number of researchers. Pearce and Wood (2019, summarising Arnot and Reay, 2007; and Taylor and Robinson, 2009 and 2013), showed the difficulty of hearing the ‘authentic voice of the student over the voice of the adults involved’ (2019, p. 119). That is, teachers and researchers have tended to mediate the voice of students through their own voice and perspective. This is the middle ground of Mitra’s three-fold classification (see Figure 18), a grey area in which the authentic voice of the student often gets lost in translation, and sometimes cannot be disentangled from the voice of the education professional. More careful use of verbatim passages can help the voice of the students to be heard in their own words; this is the approach adopted in this study.

Students whose voices ‘don’t fit the dominant discourse and academic aspirations of their schools’ tend to be excluded (McIntyre *et al.*, 2005, p. 155). There are particular challenges associated with including the voices of those not normally chosen for representational functions (Keddie, 2015; Gunter and Thomson, 2007; Cook-Sather, 2014; MacBeath, 2006; and Taylor and Robinson, 2009). Students selected by teachers for inclusion in student voice exercises often fit within – consciously or unconsciously – established ideals of “good” students (Keddie, 2015).

Given that many teachers are hesitant to consult students whose voices may offer opposing or even transgressive perspectives (Maybin, 2013), and that students themselves often prefer to adopt a more passive role, many student voice initiatives have fallen short of their transformational potential (Kehoe, 2015; Lundy, 2007; Mitra, 2006; Robinson, 2011; and Rudduck and Fielding, 2006). The institutional failure of most student voice consultations to lead ultimately to transformation may explain a lack of enthusiasm for student voice from the professional assessment community: it could be that it is simply seen as not worthwhile. Given

this, it could be argued that any study that aims to take the student voice seriously needs to engage with the thinking and research methods of the assessment community. In order to validate this approach and to demonstrate the necessary credibility, it may be prudent for such studies to use research and analytical methods that are recognized and validated within the professional assessment community, including a blend of quantitative and qualitative measures. These approaches are features of this research; they are explored more fully in Chapter 4 (Methods).

Finally, in Pearce's and Wood's evaluation, for student voice to be meaningful and transformational, it needs to be 'dialogic', where 'power is at stake' (2019, p. 120). Because the student voice needs to be heard and to become part of an exchange or conversation, this challenges traditional structures 'that position students as passive and teachers as experts and authorities' (Pearce and Wood, 2019, p. 119). These observations draw on studies by Anderson (2015); Bragg (2007); Cook-Sather (2006); Lundy (2007); and Robinson (2011).

Hall (2017, p. 188) suggests that it would be desirable to move from ideas of student 'voice' to student 'talk.' Talk involves dialogue, so there is a two-way exchange, and students' views can be both clarified and developed. Hall cites Ruddock and Fielding, where 'interaction continues with an exchange of thoughts and views' (Hall, 2017, p. 188) and the dialogue becomes an opportunity to have a say 'on things that matter to you' (Ruddock and Fielding, 2006, p. 224). These ideas of dialogic value are attractive, with their implications of gain for both students and teachers, but they are hard to implement in meaningful ways. Instead, proxies for educational value can arise.

In higher education, for example, the student voice may have 'real commercial "value" attached to it' (Naidoo and Jamieson, 2005, p. 38), with the removal of the student cap (Hillman, 2014) and rises in tuition fees. Student voice, in this respect, is in danger of taking on a political aspect, and being 'incorporated into managerialistic rhetoric' (Wisby, 2011, p. 37). A whiff of tokenism is detected here, where what the student voice says is of less interest than

the fact that it has spoken, 'as if the act of speaking is all that matters' (Thomson, 2011, p. 25). Students, in these cases, are seen as purely generators of feedback data (Groundwater-Smith and Mockler, 2016) rather than as joint constructors of any emerging understanding or knowledge about assessment (Bourke and Loveridge, 2018). Hall observes that, 'if we are not careful, evidence suggests that we become mired in "processes" and lose sight of the "voice(s)" and the opportunity for the transformational' (2017, p. 186). Iannone and Simpson point out that, in relation to their assessment preferences, 'the voice of students in the "hard-pure" sciences has rarely been heard' (2015, p. 1046).

Another proxy for real educational value (that is, something of value to both student and teacher), is the emphasis within the academic literature on matters that are of interest to the researcher, rather than perhaps of real worth or interest to students themselves, or to both parties. It is important to distinguish between these viewpoints, and for the researcher not simply to assume that what matters to the teacher also matters to the learner or, crucially, that they both see things the same way. Howard states that it is important 'that researchers analyse student perspectives of classroom instruction and learning environments since what students experience in learning may be quite different from observed or intended pedagogy' (2001, p. 133). This current study is rooted in the researcher's observations of the ways in which examination questions are used within classrooms, and aims to inform the opportunities that could be harnessed for student feedback and interaction within these learning environments.

Studies involving student voice in assessment have then, perhaps unsurprisingly, tended to focus on issues that are of more immediate interest to education professionals than to students themselves. There are other published studies, where the voices of students might have been heard but were not. Two studies illustrate this; both originate from researchers associated with Cambridge Assessment, the research arm of the examination board OCR.

Ahmed and Pollitt (2007) studied the effect of manipulating science questions from a national

science test, altering the context to make it more or less focused. The researchers manipulated questions in different permutations and tested the responses of 405 students aged 13-14. Of these students, 14 were interviewed, in pairs, immediately after completing the tests, in semi-structured interviews. The views of these 14 students informed the discussion in the article, but their voices were not heard directly: no quotations or views from the student interviews appear in the published article, but no explanation was given for this.

In the second piece of research, Crisp and Grayson (2013) applied an item difficulty modelling technique, more familiar to test developers and researchers in the USA, to multiple choice questions from UK and international A level Physics examination. Their study used the question-level results data from a cohort of 4,590 students who sat the examination in 2009. Importantly, however, the student voice was again completely absent from their evidence base: students' answers were analysed, but no-one spoke with the students themselves. The professional researchers were, therefore, the sole arbiters of difficulty in this study. It seems that an opportunity was missed here: the researchers appeared interested in the students as sources of data, but not as individuals with meaningful views.

Shaw and Crisp, in a useful summary of thinking and research on validity as applied to high-stakes assessments and examinations, commented on the benefits and some of the problems in involving students in research:

'The use of interviews with students is considered a useful activity for validation of the processes involved in answering questions. However, the current context of international A levels meant that the interviewing of students was conducted by teachers rather than the researchers, was very small scale (allowing only a small number of exam questions to be investigated in this way) and often with students for whom English was not their first language. These issues led to the data being less useful in the pilot [study on validation of general qualifications] than had been hoped' (2020, p. 9).

These potential problems and pitfalls, however, do not account for the apparent reluctance of examination boards themselves to engage with students in their research. Between 2005 and 2020, Cambridge Assessment published 29 issues of its magazine *Research Matters*. These 29

issues contained 193 articles, all on matters associated with high-stakes assessments, including validity and validation of examinations, examiner training, inter-marker reliability, and online marking practices. Only 6 of the articles involved any form of student voice. Of these, one included the views of university students (on their A level Mathematics courses), two gathered feedback from A level students, and one used data from the Longitudinal Study of Young People in England. Because these studies worked with students outside the age and qualification range that is the focus of this research, they are not discussed further here. Only two studies, fifteen years apart but both authored by the researcher Victoria Crisp – Crisp and Sweiry (2005), and Crisp and Macincka (2020) – involved the views of students taking GCSE examinations.⁵⁶ The inevitable conclusion is that UK examination boards appear not to be interested in hearing the voices of students, even though more than 600,000 of them (BBC, 2019) take their exams each year. Instead, their research activities are focused on questions of validity and validation, reliability, and accuracy.

3.4.1 Recent secondary school student voice studies – an evaluation

Four studies were reviewed that did take account of student voice in secondary schools. None of these was in a school in England, but they do present some interesting points, in terms of their methods and the degrees to which the researchers engaged with the students.

Mathematics in Tanzania

There are important cultural differences between Tanzania and the United Kingdom.

Notwithstanding these, their educational systems have many structural features in common, which enables useful comparisons to be made between the two systems. Both countries divide children's education into primary and secondary phases; both have high-stakes examinations at the end of primary school and at two similar points at the end of secondary schooling.

⁵⁶ As discussed in Section 2.1

Results of the high-stakes examinations at the end of the secondary phase determine the course of education or training to be followed in the future. As Kyaruzi *et al.* summarise,

‘The education system in Tanzania is mainly characterized by high-stake examinations which hold long-term implications for students’ lives. At the end of each instructional cycle of primary and secondary education levels, there is an external summative national examination’ (2019, p. 281)

The authors identified underperformance in mathematics among secondary school students in Tanzania. For ‘ten consecutive years (2004–2013), the majority of secondary schools students failed their mathematics national examinations’ (p. 282), which they attributed to a number of causes, including the transition from Swahili as the teaching language in primary schools to English in secondary schools; large class sizes; curriculum content overload; and lack of in-service training and professional development for teachers. Additionally, they recognised that the country’s formal programme of Continuous Assessment was in fact another form of summative assessment, without the educational benefits that a programme of formative assessment might provide. They therefore set out to discover students’ attitudes to formative assessment in mathematics, where such feedback was provided.

A large-scale study was undertaken, including collecting data from 2,767 students across 48 secondary schools, evenly divided between urban and rural settings. The authors deployed a mixed-method research approach, with quantitative (survey) and qualitative (focus group discussions) methods. A conceptual model for their study was first developed. A correlational survey design ‘using a two-step process (Anderson and Gerbing, 1988) of first establishing robust measurement models for each construct [using previously validated questionnaire items], followed by a structural equation model linking the constructs as outlined in the conceptual model’ (p. 283). Content analysis of qualitative data from the focus group discussions was then linked and correlated to the survey findings. The researchers noted the important caveats that the cross-sectional nature of the survey made it impossible to draw strong causal conclusions, even where there appeared to be significantly positive correlations

between different measures of the same construct, and that there was no independent observation of teacher practices to see if these were consistent with the students' reports.

Drama in Australia

Hogan (2018) investigated how secondary school students in year 10 in three schools in Queensland, Australia, described their experiences of teacher feedback in drama lessons. This was a relatively small-scale study, although obviously time consuming, with 57 students participating in classroom observations, a smaller subgroup of 37 students involved in focus group discussions, and individual interviews with 24 of these students. Hogan used a multiple-case study methodology (cited Stake, 2006). The purpose of this practitioner research is clearly to improve teachers' understanding of, and effectiveness in using, formative feedback to students. Methodological weaknesses are, however, evident in this study, some of which may be considered to encroach on the validity of the findings. The researcher relied on her field notes and analytical memos from classroom observations, and verbatim transcriptions of focus group discussions and individual interviews. There were no student surveys or questionnaires, and no quantitative data was collected. No conceptual model is presented. Analysis is entirely thematic, and no details are given about any coding techniques used to sort the qualitative data. Conclusions are presented as 'emergent findings', backed up by plentiful quotations in students' own words, but it is hard to evaluate to what extent they objectively represent the balance of students' views. The findings are discursive, leading to a number of specific recommendations. No triangulation was made with teachers' own views of how their feedback was received.

English and mathematics in Australia

Van der Kleij (2019) investigated similarities and differences in feedback perceptions among teacher and students, specifically in relation to secondary school learning in English and mathematics. The study also explored the association between individual student characteristics and students' feedback perceptions. Survey data was collected from 59

teachers and 186 students across five Australian schools. The ages of students ranged from 12 to 16. Feedback quality was perceived more positively by teachers than students. Students' self-reported levels of self-efficacy, intrinsic values and self-regulation predicted their perceptions of feedback quality.

This is a study in two parts, with an admirably clear and simple structure. A conceptual model was developed of the student perspective in the feedback process. Study 1 comprised two surveys, one for teachers and one for students, using an existing self-report survey of feedback practice, translated from Norwegian and modified to suit the research purpose and Australian context. Parallel questions in the teacher and student surveys allowed straightforward comparisons of perceptions to be made between the two groups. Two different dimensions of student and teacher feedback were investigated. The survey made use of a four-point Likert scale without a neutral response option. Surveys were conducted online, separately for English and mathematics. Factor analyses and Cronbach's alpha reliability analyses were deployed to validate the survey instrument. Data from the Likert scale items were coded numerically to allow for quantitative analysis. Sophisticated statistical techniques were deployed to enable valid comparisons to be made between populations that exhibited different properties. Although teachers' responses showed a nearly normal distribution, the 'distribution of scaled student scores differed significantly from a normal distribution' in both subjects, and so 'non-parametric Mann-Whitney U-tests were used when comparing teacher and student feedback perceptions. Wilcoxon matched-pair signed rank tests were used for within-group comparisons' (p. 179). Study 2 compared students' perceptions of feedback against a range of self-reported psychological characteristics. Similar statistical techniques were used for the second part of the study. Some additional free-text responses were quoted to give additional qualitative perspectives and explanation to the findings.

Barrance and Elwood (2018) used student voice perspectives not to critique or evaluate pedagogy, but to give an account of students' actual experiences and perceptions on the 'debate around assessment policy reform within the context of devolved government arrangements for assessment and qualifications' (p. 253). Citing Bragg (2007) and Thompson (2011), the authors reflect on 'shortcomings associated with the work around student voice', particularly in 'the tendency to perceive some students' voices to be representative of all others' (p. 257). Drawing on a data set previously compiled from a large mixed-methods research project (Ellwood *et al.*, 2017), the study used questionnaire survey data from 1,600 GCSE students across Wales (901) and Northern Ireland (699), and 20 focus groups each of 5-10 participants. An innovative aspect of the study was that it involved students as advisors to the research in the development of the research instrument. The analytical method is worth quoting verbatim, for its succinct summary of the mixed methods technique and for the way it includes student researchers in the analysis:

'Quantitative data from the survey was analysed using SPSS to identify relationships between key variables and patterns of responses of interest to the overall aims. Qualitative data was (i) coded by hand if it came from the open-ended questions on the survey and (ii) transcribed and coded using MAXQDA if it was from the focus groups. In analysing the data, the advisory group members [the students themselves] were also involved in looking at anonymized extracts of qualitative data to help code and arrange into themes' (p. 264).

(SPSS is quantitative analysis software; MAXQDA is qualitative analysis software, offering mixed methods, statistical and quantitative content analysis.) By involving the students themselves in the analysis of the data, the authors of the study consider that the quality of the analysis was enhanced, as also was the credibility of the findings with young people themselves. This is an interesting and rare example (outside the sphere of Youth Participatory Action Research) of ensuring that the authentic voice of the students is heard at different points in the design, data gathering, analysis and presentation stages of the study. Arguably, this is as important a contribution to the literature on student voice as the findings

themselves, as it shows 'students as having the capacity to express valid opinions on complex assessment issues' (p. 266). As for the outcome of the research, students tended to agree with the substance of their governments' assessment reforms, even if not with their expressed reasons for them. Similar approaches, both in the involvement of student researchers and the quantitative and qualitative analysis methods used, feature in Barrance's 2019 study evaluating students' views on the fairness of internal assessment in GCSE.

From these studies using student voice in secondary schools come some valuable insights into suitable methods for use in this study. The use of a survey questionnaire, perhaps using a Likert-type scale, to capture the views of a large number of students (if possible) appears to give a thorough grounding, on which smaller, more intense and focused group conversations can then be based to give a wealth and balance of students' views. Teachers' voices can feature occasionally for context and for validation of some of the processes and learning experiences that students describe. The more interesting studies (for this researcher) were the ones that got closest to hearing, presenting and interpreting the authentic voices of students. These are aims and features that will be taken into the next chapter, Methods.

3.4.2 Student Voice in Higher Education and in Primary Schools

There is a relatively large number of published research studies involving student voice in connection with assessment in higher education. The focus here is therefore restricted to more recently published studies. Using search tools on ERIC, ProQuest and Web of Science, Sun *et al.* (2022) identified 373 articles published between 2011 and 2022 that sought to draw on student experience of assessment and feedback in higher education. Having filtered these studies further, to include only studies that were published in English, peer reviewed, and included primary research, Sun *et al.* carried out a systematic review of the methodologies and aims of the remaining 38 studies. Synthesising the findings of the studies, there are many reasons given for higher education institutions to engage with student voice, ranging from a genuine interest in learning about students' experiences and incorporating their views to

improve university teaching (Canning, 2017) to, more formally, fulfilling regulatory requirements to conduct student voice surveys (for example, Matthews and Dollinger, 2023).

The large majority of these studies with higher education students were carried out in, or with students from, developed countries, with the UK, Australia and the USA accounting for the most. Research methodologies in the review period 2011 to 2022 showed an increase in the use of mixed methods and quantitative studies, with larger-scale samples of students.

Qualitative methods were less often seen over this period, and they were always with smaller sample sizes. Although some authors were cautious about the extent and effectiveness of student participation (Mendes and Hammett, 2023, p. 164, wondered about a new 'tyranny of participation'), many appeared enthusiastic about the opportunities for collaboration: Cook-Sather (2020, p. 898) wrote that drawing on student voices 'supports meaningful dialogue that breaks down traditional barriers between instructors and students⁵⁷.' This laudable aim was reflected to some extent in the aims and purposes to which the student voice exercises were put: Sun *et al.* found that just over one third (37%) of the studies aimed to improve student experience; the same proportion aimed to improve teachers' assessment practices; developing the teaching and learning environment was the next most common aim. It is evident that it is far more common practice to involve students in giving feedback around assessment in higher education than in other phases of education, but it is not evident that these voices are always heard in ways that are effective in improving the quality of assessments.

Many fewer published studies were found that involved student voice in primary schools in connection with learning and assessment. The period under review was extended back to 2005 in order to accommodate studies included in Pearce's and Wood's 2019 cross-phase systematic review. Robinson (2014), building on work previously undertaken by Robinson and

⁵⁷ To contextualise this rosy view, the same author rather startlingly revealed that students in secondary schools are now regarded as 'equal partners in the evaluation of teaching and learning' (Cook-Sather, 2018, p. 18), which might be surprising news to many students, teachers, and teaching unions.

Fielding (2007, 2010), reported the increasing number of initiatives in the United Kingdom (albeit from a very low base) aimed at eliciting the views of children. In contrast to the practice in higher education, almost all the primary school studies reviewed by Robinson were small or medium-sized; methodologies usually included surveys and semi-structured interviews. In matters related to this current study, Hopkins (2008) reported that some primary-aged pupils made a link between challenge and their motivation, enjoying the struggle and stretch of more difficult work. Chamberlain *et al.* (2011) reported that many upper primary aged pupils felt unhappy and weary about the amount of pressure teachers put on them in preparation for national assessments (SATs). Only one study (Wellcome Trust, 2010) focused specifically on assessment in primary schools, and this was within science. This study involved 1000 children aged 10-12 in England and Wales. Primary pupils in this study voiced concerns about the negative impact of assessment pressure. However, in the context of the imminent abolition of Science SATs⁵⁸ by the UK Government in 2009, pupils were mostly against the abolition, voicing their views that it might lead to them not learning as much about science, not being aware of their levels of achievement in science, and that this might also lead to science becoming less important in schools. This perspective from primary pupils is interesting – with its resonance of ‘washback’ from assessment into teaching, and that ‘what is valued tends to be what is assessed’ (Fautley, 2015, p. 513) – although it may be noted that it comes from a report commissioned by a charitable trust set up to promote science education.

In terms of the timing and format of classroom assessments, Robinson (2014, p. 15) found that children strongly favoured being tested just after they had completed a topic, rather than later in the year, and that they disliked the ‘traditional pen-and-paper, sitting at a desk approach,’ preferring more active and ‘fun-type’ assessment styles: presentations, investigations, research, group work and project-based assignments. This understandable pupil preference for active assessment formats – which secondary school teachers will also recognise – is

⁵⁸ In English schools: science SATs had already previously been abolished in Welsh schools, where decision making about education is part of the powers devolved to the Welsh government.

unfortunately at odds with the continuing dominance of closed-book invigilated terminal examinations that await these pupils in their secondary and further/higher education.

Bringing Robinson's 2014 study up to date by extending her search methods up to 2023, only two published studies were found⁵⁹ that involved primary pupils in talking about assessment in the context of teaching and learning. Florian and Beaton (2017) used video recordings of lessons, semi-structured interviews and discussions with small numbers of pupils and teachers in one English primary school – narrowing down to one class – to focus on how teachers can improve the quality of formative assessment by listening to pupils' self-assessments of their learning needs and next steps. Their study makes extensive use of verbatim quotations to bring the pupils' voices vividly into the discussion. Leenknecht and Prins (2018) carried out an experimental study into formative feedback in one school in the Netherlands, with 95 pupils randomly assigned to either a treatment or a control group. Pupils in the treatment group were involved in setting their own assessment criteria and choosing the style of feedback for their formative assessment. Compared with the control group, pupils in the treatment group appeared much more engaged and participated more fully and confidently in listening to and acting on formative feedback. No students' words were quoted in this study.

Although limited in number, studies that use student voice in connection with assessment and learning with primary pupils have shown some commonality: with the exception of the Wellcome Trust (2010) report into science education, they have been small in scale, using qualitative methods and carried out by external researchers. These studies have indicated that primary school students are aware of and can be engaged in issues regarding assessment and learning. They have valuable opinions that, on occasion, have helped teachers to refine their formative assessment practices. In these studies, primary pupils show that many of the same

⁵⁹ Using Google Scholar, ERIC and university library search engines, using filters for combinations of *student/pupil voice, assessment, learning, primary and school*.

thoughts about motivation, stress and the styles of assessment that will emerge later in secondary school are already present in their younger years.

3.5 Literature Review Conclusion

In concluding this extensive review of literature that forms the backdrop to the present study, it will be useful to look again at the research question:

- **How do students experience and comprehend demand and difficulty in GCSE mathematics examination questions?**

At the beginning of this chapter, a diagram was given (Figure 4), showing how the different theoretical models fit together in this study. The passages that follow present a very short summary of the four areas of literature reviewed, an indication of how the understanding gained in each one relates to the research question, and some explanation as to how they lay the ground for the study that follows.

Working memory and cognitive load theory

From the survey of literature on working memory and cognitive load theory, it was established that the working memory of a student creates an effective bottle neck for cognitive activity, and that examination stress may diminish this further. Sweller's cognitive load theory, linking with working memory theory, shows how a student commits knowledge and understanding to their long-term memory, via the rehearsal of this knowledge – perhaps through attempting to answer examination-style questions – and the acquisition of schema. Cognitive load theory, although it has limitations, provides guidelines for the design of effective instruction. In the study that follows, it will be important to find out whether students comprehend in their own experience the link between previously learned knowledge and their ability to apply this to the new demands of an unseen examination question, and whether they understand the role that environmental factors such as exam stress and a shortage of time can play in their ability to manage these demands.

Taxonomies of learning

When answering a question, the student recalls pre-learned knowledge and understanding, and applies these to the demands of the question. From the engagement with literature on taxonomies of learning, it was established that the role of knowledge in answering questions is complex, and occurs at different levels. It is in the centrality of the role of 'knowledge,' fuelling all levels of processing in the cognitive system, that Marzano's and Kendall's New Taxonomy (2007) appears more adaptable, relevant and applicable than Bloom's Taxonomy in explaining how students cope with and meet demands in examination questions. Its multi-dimensional approach, avoiding a simply hierarchical structure, makes it more flexible than Bloom's in describing complex learning processes. Bloom's Taxonomy is almost ubiquitous in teacher education, however, and so it provides a useful starting point for working with teachers and students in understanding demand and difficulty; but working with Marzano's and Kendall's New Taxonomy is more likely to provide a firm basis for interpreting and presenting students' experiences and comprehension of demand and difficulty in examination questions. To give a relative measure of the demands imposed by examination questions, scales of complexity, resources, abstractness and response strategies (CRAS scales) were devised by Pollitt *et al.* (2007). These will be used to gain a sense of the demands of the examination questions chosen for the study, so that students' experience and understanding of these same questions can be contextualised. The Noel Burch Competency Model enables the effectiveness of the students' conscious and unconscious cognitive processes to be usefully classified, described and compared.

Validity

Thinking on validity further informed the purpose of this study: validity is a complex concept, and its meaning has been shaped and developed over time by different concerns around assessment. This study takes what Newton (2012) has described as the 'consensus' view of construct validity, that it is the interpretation of the result that is valid (or not valid), and not

the test itself. There will be many possible interpretations of the single reported result of a GCSE mathematics examination, and it is by no means clear that the examination system can support them all. Spalding (2009) made the clearest link between the ways in which examination questions function and the validity of the inferences that can be drawn from the results of examinations.

Having explained that the first duty of examiners is to construct examination questions and papers that enable them to 'deliver scores for students that show as accurately as possible how much and how well they have learned,' Pollitt and Ahmed (2011, p. 260) outlined three threats to validity. The one that is of most relevance to this study is the idea that there may be features of the question that elicit behaviours from students that are different from those expected by examiners. In terms of construct validity, the threat here is that, if certain features of one or more questions lead to students misinterpreting the demands, so that their answers are not a reliable indicator of their expertise and prior knowledge in respect to the questions, the marks produced from the examination process will lose their value as a straightforward indicator of the relative expertise of the student. From this, it follows that interpreters of the examination grades obtained would not be able to make valid inferences about the relative merits of the students. Examiners are only one of many groups of interpreters of examination results, however, and the implications of the grades they produce are arguably not as significant to them as they are to the students themselves and to future users such as employers and further education providers. However, it is examiners who must bear the burden of the responsibility to ensure that they understand how well their questions perform in practice, in order that this construct validity can be as secure as possible. In this study, a thorough understanding of construct validity therefore prepares the ground for investigating how students experience and articulate their response to any unpredictable demand features of examination questions.

Student voice

Finally, from this chapter's review of published literature about student voice in research, it was established that there have been few immersive studies using student voice in secondary schools in relation to learning and assessment, and that the existent studies have often tended to be what Hall (2017) labelled 'tokenistic' rather than 'transformational.' There has been more optimism around the use of student voice in higher education to improve their assessment experience, matching the shift for students from passive to more active participants in this sector. I approach my topic in this study with a genuine desire to collaborate with young people in the secondary phase of their education. Following the stance of Mitra (2018), I aim to hear and present the authentic voices of students and, following Apple *et al.* (2009), to harness and interpret their views in an effort to improve both the quality of standardised assessment and the quality of advice and training that can be given to teachers and students. In the next chapter, I set out in detail the methods – and explain their rationales – through which I intend to hear and present these student voices.

Chapter 4: Methods – Planning my immersive journey

In this chapter I plan my journey in more detail. Like Robert Frost's traveller, where "Two roads diverged in a wood, and I – I took the one less traveled by"⁶⁰, I opted for qualitative research methods as the most suitable route for my journey.

This chapter sets out, and provides academic justifications for, the methodological approaches adopted within this study. In addition it explores the philosophical positions taken, the design of the pilot and main studies, discusses factors that influenced the choice of questions, and also comments of the methods of analysis deployed within this study. Finally, ethical issues pertaining to, amongst other issues, the collection and use of data within this study are noted.

Waring (2021a, p. 17) observes that the question 'what data collection techniques or procedures should be used?' is simply answered: 'it is those techniques and procedures which allow the researcher to gather data that are appropriate to answer the research questions'.

The research question addressed by this study is:

- **How do students experience and comprehend demand and difficulty in GCSE mathematics examination questions?**

In order to evaluate the lived experience of students and to collect and present a sense of their comprehension of examination questions, research methods have been chosen that are sufficiently responsive to capture the sense and meaning of students' spoken and written words. To value and weigh what students say and write, qualitative methods are most appropriate. These methods enable the researcher to collect and analyse non-numerical data, to gather in-depth insights into the problematic concepts of demand and difficulty in natural rather than experimental settings, and to generate new ideas for research (Agius, 2013).

Qualitative methods are used in this study to gain an understanding of the complex experiences, perceptions and behaviour of the students who are the principal stakeholders in

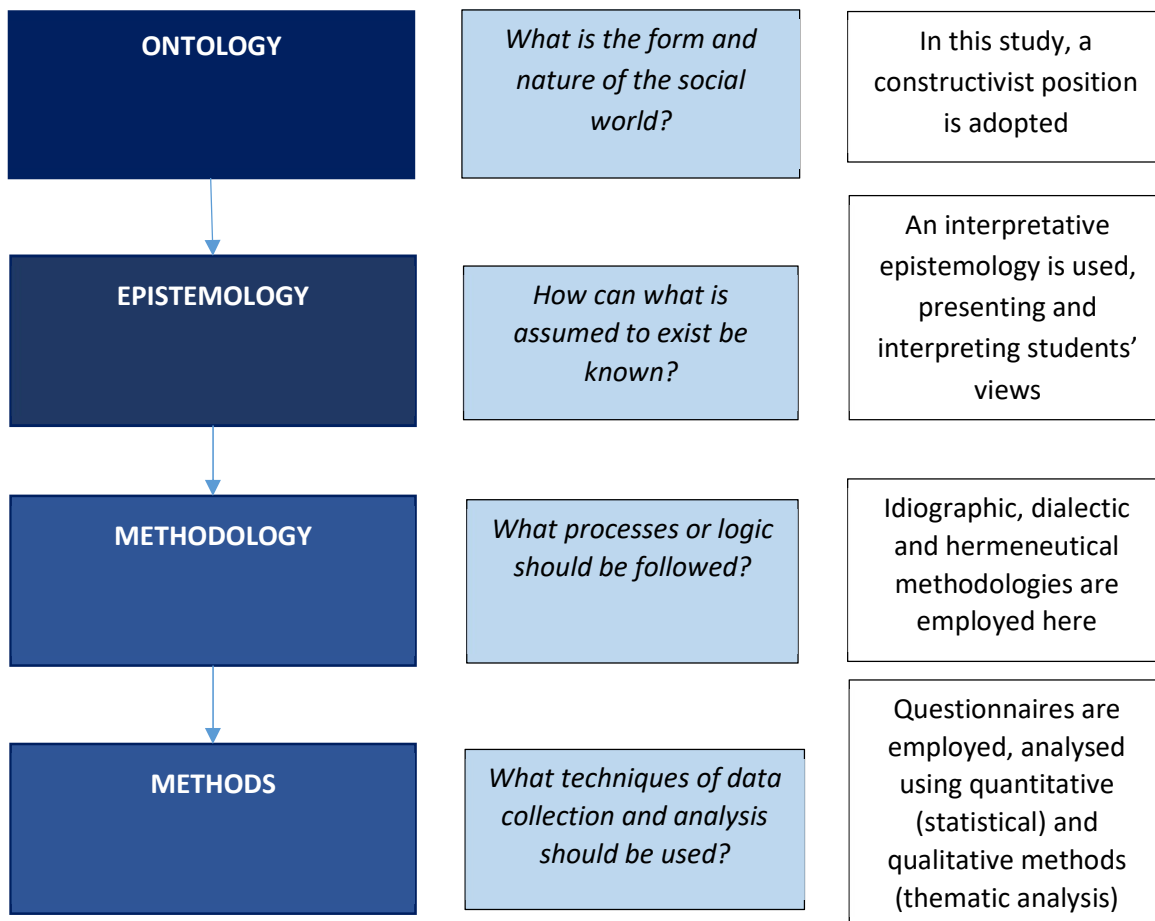
⁶⁰ Robert Frost – *The Road not Taken*, 1916

examinations, and to investigate the meanings attached to these experiences and perceptions. Straightforward statistical measures are used to report numbers, averages and trends.

4.1 Philosophical location of the current study

Grix (2002 and 2018, as cited in Waring, 2021a, p. 15) notes that there are four “building blocks” of research: ontology, epistemology, methodology and methods. To assist in the philosophical positioning of this thesis, these four building blocks are presented in a diagrammatic form in Figure 19.

Figure 19 - Relationship between Ontology, Epistemology and Methods in this Study



Source: (left-hand column) Waring (2021a, p. 16); (middle column) condensed from Waring; (right hand column) author's own.

Ontology is concerned with the form and nature of the social world: it is 'the nature of reality' (Lincoln and Guba, 1985, p. 37) and 'the study of being' (Crotty, 1998, p. 10). Within the present study, a constructivist approach was taken, because the lived and perceived reality of demand and difficulty within GCSE examination questions can vary considerably from one individual student to another; this study attempts to uncover the range of meanings that students bring to their discussions of these questions. This constructivist stance makes particular sense in the present study, because an examination question that appears easy to one student may appear difficult to another, depending on their relative levels of skill, knowledge, and understanding of the subject or topic; their confidence and self-efficacy; and possibly other circumstances operating at the time the student tackles the question. The varied experiences of students have been analysed and distilled, constructing themes from the responses students gave, and letting their diverse voices be heard within this contribution to research literature on examinations. Ontology and epistemology are closely linked in Waring's model (as shown in Figure 19). Indeed, Crotty notes that an ontological stance implies a particular epistemological stance and vice versa. Epistemology is 'how we know what we know' (Crotty, 1998, p. 8). Establishing this study's position on epistemology involves a consideration of the relationship between knowledge, truth, belief, reason, evidence, and reliability.⁶¹ This study identifies itself as having an 'interpretative' epistemology, because it intends to 'acknowledge and include the perspective and voices of the individuals' involved in the study (Waring, 2021b, p. 120). Operating under a constructivist ontology, an interpretative epistemology 'does not see direct knowledge as possible; it is the accounts and observations of the world that provide indirect indications of phenomena,' (Waring, 2021a, p. 16). In this way, understanding is defined and developed through a process of exploration and interpretation.

In adopting a constructivist/interpretivist approach, this study aims to understand students' perspectives on demand and difficulty by seeing the world of GCSE examination questions

⁶¹ Epistemology, as defined by Sheffield University Philosophy Department in <https://www.sheffield.ac.uk/philosophy/research/themes/epistemology> accessed 06.07.2023

through the students' own eyes, using their words where possible to articulate their experience and understanding. In this way, the researcher's position is relativist, transactional and subjectivist (in accordance with the views of Guba and Lincoln, 1998), because the researcher and the social world impact on each other. Within this study, I, as the researcher, am both a cultural member, immersed in the culture which the research subjects – the students – inhabit, and I am also a cultural commentator. As a member of the educational and social culture from which the survey responses are drawn, I am able to contextualise and interpret student responses, and through so doing I become part of the dialogue between theory and practice. I am a teacher and a headteacher, and through these positions I have a thorough and immersive understanding of the cultural and educational context of the respondents, all of whom are rooted in the schools and the area in which I live and work. But in analysing and reporting the experiences of students, as author I also become a cultural commentator, providing contextualised commentary and analysis on the views and responses I interpret. Finally, an important part of any rigorous approach to epistemology is also to include a healthy pinch of scepticism (Waring, 2021a): a realisation that even well-founded interpretations are by their nature subjective, and that researchers can never truly know the reality behind the appearances of the data. Taking time to collect and analyse the data in this study – space, time and distance that were partly enforced by the COVID-19 pandemic – enabled me to question and revisit my interpretations and assumptions.

The methodological assumptions made in this study reflect and relate to the assumptions made under the ontological and epistemological “building bricks.” Under a constructivist ontology and an interpretative epistemology, methodologies tend to be more idiographic, dialectic, and hermeneutical (Waring, 2021a, p. 16). In the present study, an idiographic approach was taken through the views of individual students providing the starting point data for the research. These individual opinions and experiences were investigated through logical and reasoned argument (that is, dialectic), looking for the “meaning” of the student experience by seeking contextualised explanations and further evidence, supporting or

contradictory. As qualitative student responses were considered, read and re-read, themes were constructed, which led to a hermeneutical (interpretative) study of the language used by students and the meaning they were seeking to convey; these aspects are, in terms of practical results, subsequently discussed in Chapters 5, 6 and 7.

Methods of data gathering and analysis need to match and support the research question (as restated at the outset of this chapter). The constructivist approach lends itself particularly well to a qualitative approach, gathering data through questionnaires and focus group interviews, and bringing thematic analytical methods to bear in order to capture and explore the richness and variety of student experience. The context for the qualitative discussions is also important, however, and this brought some elements of a mixed methods approach into this study. In order to focus the discussion with students on their experiences and comprehension of demand and difficulty in examination questions in GCSE mathematics, it was necessary for students to engage with a sample of those questions. As a teacher, my experience suggested that it was valuable for students to attempt the mathematics questions, in some approximation to an examination environment, if possible. I also wanted to investigate whether and how students' perceptions of difficulty matched up with their actual experience of difficulty. To investigate this, I needed students to answer the questions as well as to offer opinions about them. Consequently, paper and online surveys were used to gather the views of students.

Some of the survey questions – about the grading of relative difficulty of examination questions, for instance – gave rise to simple “scale” responses (in this case a Likert-type scale), whilst other questions gave rise to free text responses so that students could express the detail and range of their thoughts. Simple statistical methods were used to investigate response patterns to the mathematical questions and estimations of difficulty. Semi-structured focus group interviews were carried out to investigate further some of the themes and

approaches from the questionnaires. Reflexive thematic analysis was adopted as the principal method for investigating the qualitative data.

4.2 Ensuring validity

When collecting and analysing quantitative data, validity relates to positivistic principles: knowledge is derived from empirical evidence, observed through sensory experience (in this case, what it looks like) and interpreted by means of reason and logic. Validity may therefore be maximised through ensuring controllability, replicability, predictability, randomisation of samples, objectivity and observability (derived from Cohen, *et al.*, 2018, pp. 246-247). The parameters are different in qualitative research and, consequently, different approaches to validity are needed. Winter (2000) suggested that validity in qualitative data might be addressed through the honesty, depth, richness, and scope of data recorded, the range of participants approached, the extent of triangulation undertaken, and the objectivity and disinterest of the researcher. Qualitative research necessarily involves one human being (the researcher) interacting with other human beings (the subjects), all of whom bring possibilities of human error to their activities. In this already interpreted world, the researcher engages in a doubly hermeneutic exercise (Giddens, 1979) to understand other peoples' understanding of the world.

Within this study, I am the author, and I am both the researcher and a headteacher in a school. Within the pilot study, I engaged with students who are similar ages to one another but who have different experiences and different abilities from one another, who are taught by different teachers and who attend a range of similar but subtly different comprehensive schools. The period of data collection for the pilot study was during a global pandemic in which schooling was substantially disrupted. The main study was conducted in person, in the school of which I am the headteacher, through class surveys and focus group interviews with volunteers from two of the four classes surveyed. It is inevitable that there will be multiple

layers of interpretation operating. Notions of controllability, replicability, observability and so on, therefore operate in very different ways from those encountered in, for instance, a large-scale randomized control trial.

According to Cohen *et al.*, (2018, p. 247), qualitative research 'abides by principles of validity which differ in many respects from those of quantitative methods.' Maxwell (1992), suggested that 'understanding' is a more suitable goal (and term) than 'validity' for the qualitative researcher. Researchers, he argued, are part of the world they are researching: they cannot – and perhaps should not aim to – be completely detached from this world, as the tenets of true objectivity would dictate. But neither should researchers import their own biases: they should acknowledge that other people's perspectives are as valid as their own, and they should attempt to uncover and understand them. Agar, cited in Silverman (1993), claimed that qualitative data collection necessitates an intensely personal involvement of the researcher, and that this involvement, and the in-depth responses of individuals secure sufficient levels of reliability and validity. Hammersley (1992) and Silverman (1993) disagreed, however, stating that the close involvement of the researcher and the reporting of detailed responses are, in themselves, insufficient grounds for assuming either validity or reliability. In order to ensure the validity of the inferences made from the data collected for this study, great care was taken in the design of the research instruments and the application of the methods of analysis.

Lincoln and Guba (1985, p. 219), as well as Cohen, *et al.* (2018, p. 248), have suggested several steps (including researcher immersion in the field, and taking due account of all data) that researchers can take to increase credibility. In a similar manner, Onwuegbuzie and Leech (2006, pp. 239-46) set out many steps that can be taken to ensure the conditions for validity in qualitative research are as strong as possible, whilst principles for ensuring the validation of qualitative studies have been collated and discussed by Cohen, *et al.* (2018). As Tables 2 and 3 illustrate respectively, by employing these steps and principles, this study has sought to ensure the validity of its conclusions.

Table 2 - Steps to Ensure Validity in Qualitative Research in the Current Study

Steps the researcher can take to ensure validity in qualitative research	Evidence for these steps in the current study
Prolonged engagement in the field, to gather data that is sufficiently rich	The researcher has worked as a secondary school teacher for 25 years, and in senior leadership for 10 years. He has undertaken three previous studies in assessment with secondary school students (one of which has been published in a peer-reviewed journal), using a mixture of quantitative and qualitative methods. The present study draws on data gathered in 2021 and 2022, allowing sufficient maturation in the field
Persistent observation, to identify key relevant issues and separate these from comparative irrelevancies	The present study is the result and culmination of a number of studies over many years of part-time study (2015-2023), and of discussions with teacher colleagues, supervisors and other researchers
Triangulation	Students' perspectives from the two studies were compared and contrasted. They were also triangulated with the researcher's own classroom experience and other published studies, where these are available
Leaving an audit trail of documentation and records, including process notes on how the research is progressing	All the student responses generated in this study, whether on paper or electronic, have been preserved and can be made available for examination and audit. Notes on the process were kept in notebooks and as part of supervision records, and can be made available
Member checking and respondent validation	All students who took part in the study were recruited through their teachers, with the consent of their heads of department and their headteachers. Names and basic demographic data were collected, as part of the validation and consent-checking activities at the beginning of every survey. Names were removed before data analysis took place and were replaced by codes, to reduce possible researcher bias. Any spurious names or anonymous responses were discarded
Weighting evidence, ensuring more attention is paid to high-quality longer engagement with data	Greater attention was given in analysis and discussion to longer and more detailed responses, and to sustained conversations undertaken through focus group interviews, some of which generated some of the most interesting data
Checking for representativeness, avoiding generalising of unsupported findings	Where particular responses were isolated and did not form part of a pattern, account was taken of them (as exceptions) but they were not extrapolated and generalised. During the reflexive and immersive process of thematic analysis, unrepresentative and unsupported responses were filtered out. Some anomalous findings were reported, in the interests of balance, but they were not allowed to outweigh more representative themes

Steps the researcher can take to ensure validity in qualitative research	Evidence for these steps in the current study
Clarifying researcher bias, avoiding interference or influence of the researcher's own personal characteristics; peer debriefing of conduct and findings	Research for the pilot study was carried out online and there were few respondents from the researcher's own school, so the contact between researcher and respondents was minimal. There was therefore very little opportunity for any respondent to be influenced by characteristics of a researcher they did not meet, either online or in person. For the main study, all student participants were from the researcher's own school. There was some engagement between the researcher and the research subjects. The strengths and possible weaknesses of this immersive approach are discussed in this chapter. Methods and data were discussed with supervisors and with professional colleagues, to allow alternative perspectives and challenge to be considered
Theoretical sampling, following the data where they lead, rather than leading the data	The study adopted a constructivist and interpretivist approach, because there was so little previous research in this area and there were no <i>a priori</i> hypotheses to adopt
Making contrasts, for example, between groups and sites	The pilot (online) study was divided into two groups, and students at five schools were allocated to one of the groups. This took place by schools, for ease of administration. In the analysis and discussion section, frequent comparisons and contrasts between the two groups have been explored. In the main study, all students were from the same school, and the method of data collection was in-person, through questionnaires and focus group interviews. Contrasts between the two studies are brought out in the discussion section
Checking meaning of outliers, rather than ignoring or eliminating them	After the initial data cleansing, to remove spurious data, no responses were eliminated or ignored. Where variant readings and responses were given, these are acknowledged and explored in the discussion
Using extreme cases, finding out what is missing from the majority; following up surprises	There were few extreme responses, but anomalous and alternative responses were investigated, reported and discussed. In some cases, these provide illuminating viewpoints and throw other responses into relief
Replicating findings	Owing to the nature of the study, it was not possible to use this technique. This will be discussed in the section on limitations and recommendations for further study
Referential adequacy, ensuring that findings are well referenced to benchmark or other significant literature	A wide range of literature was used for reference, including theoretical writing, research studies, literature produced for and by examiners, and professional literature. Details can be found in the reference section of this dissertation

Steps the researcher can take to ensure validity in qualitative research	Evidence for these steps in the current study
Structural relationships, looking for consistency between the findings, with each other and with literature	This was a feature of one of the main methods of qualitative analysis, coding responses and looking for patterns, consistency, agreement and disagreement, and matching this with relevant literature, where possible
Rich and 'thick description,' providing detail to support and corroborate findings	Extensive verbatim quotation is used in the analysis and discussion sections, providing access and insight for the reader to the authentic voices of the students

Source: (left-hand column) Adapted from Onwuegbuzie and Leech (2006), and Cohen, *et al.* (2018, p. 249); (right-hand column) Author's own notes on evidence in the current study.

Table 3 - Validation Principles for Qualitative Study, Related to the present Study

Validation principles for qualitative study	Reflections from the present study:
The natural setting is the principal source of data; data are socially situated, and socially and culturally saturated	The two qualitative studies were both carried out in the context of state-funded secondary schools in England; all participants were students in their GCSE years (ages 14-16)
The researcher is part of the researched world; the researcher (rather than the research tool) is the key instrument of research	The researcher is a headteacher (of one of the schools in the study), using his own experience of teaching the subject in question to try to elicit a greater understanding of the students' views and knowledge
Data are context-bound, and descriptive; data may be characterised by 'thick description'	The school and examination contexts for the two studies have been discussed in detail, as part of the study. Every effort has been made to reflect the context of students' individual comments
Data are analysed inductively rather than using <i>a priori</i> categories	A reflexive thematic analysis approach was adopted, with an interpretative epistemology, so that levels of abstraction were generated from the inductive analysis of the students' responses; no <i>a priori</i> categories were applied (indeed, none existed)
Data are presented in terms of the respondents rather than the researchers; seeing and reporting the situation through the eyes of participants; catching meaning and intention are essential	Wherever possible, students' comments have been given in their own words, including where these are ungrammatical or mis-spelled, in order to capture the authentic voice of the participants. Further details have been added, where additional context in relation to individual participants helps to give the bigger picture
Validation of respondents is important	Respondents were all approached through their teachers, having first gained the consent of the headteachers of the individual schools. Students all had the opportunity to give or withhold consent. The responses of consenting students only were analysed. All the records, electronic and paper, have been retained, so a suitable trail of evidence remains that can be audited if necessary to verify the participants and their views

Source: (left-hand column) Validation principles derived from Cohen *et al.*'s synthesis (2018, p. 246); (right-hand column) Author's own notes in relation to current study.

4.3 Ethical considerations

This study follows the guidance published by the British Educational Research Association (BERA, 2018). Two proposals for research were submitted to the University of Durham School of Education's ethics committee, and approval was given.

In order to protect the privacy of students and the confidentiality of data, suitable features were designed into the research methods. Data codes (group / student number / gender) were assigned to the responses before analysis, following approaches described by Saldaña (2013, pp. 25-30). Individual students are therefore not identifiable from the data codes assigned. In order to respect the autonomy of students participating in this research, and to enable them to make informed decisions about their participation in the research, students were informed about the purposes and potential benefits of the research, in accordance with general principles outlined by Powell *et al.* (2012). Because the student participants in this study were under the age of 16, they were not deemed able to give informed consent for their participation in the study. Instead, their headteachers and their heads of department were approached and they gave consent. Nonetheless, students were also asked for their consent, in line with this study's philosophy of valuing students and listening to their voices. In the case of the researcher's own school, consent was sought and obtained from the school's Chair of Governors. Because the risks were deemed to be low, and the research activities were very much within the normal run of classroom activities (Esbensen *et al.*, 2008), no attempt was made to obtain active parental consent. For the pilot study, suitable information was provided online to students at the start of the questionnaire (Appendix A). An information sheet and consent form was provided to students in the main study (Appendix B). Main study students completed the questionnaire in the context of a class lesson, so it was not suitable for them to decline to participate in the lesson activity, but they were able to indicate on the consent form whether they wished for their responses to be included in the research study. Pilot study

students chose whether to give their consent before the online form directed them to the questionnaire. Those who did not give consent ended their participation at this point: they were not referred to the pilot study questionnaire. 61 students out of a total of 192 (31.8%) chose not to give consent and therefore did not complete the pilot study questionnaire. In the main study, all of the 97 students gave their consent. Students in both the pilot and main studies were informed of their freedom to withdraw from the study at any time in accordance with the approach advanced by Powell *et al.* (2012). At the time of submission of this thesis, no students had asked to withdraw.

4.4 Questionnaire design

Since the focus of this study is demand and difficulty in examination questions, and how students comprehend and relate to these concepts, the choice of which particular examination questions to use for research purposes, out of the hundreds available, was an important one. It was decided to focus on questions in mathematics, partly to limit the subjectivity associated with (particularly the marking of) more discursive subjects such as history, and partly because this was among the teaching specialisms of the researcher, meaning that greater levels of expertise and a more rounded professional perspective could be brought to bear in the discussion of results.

For the pilot and main studies, GCSE mathematics questions were selected. Seven questions were selected from past papers. Six of these were used for the pilot study, and one additional question was added to the main study for comparison purposes. For consistency, the first six questions were taken from the same examination board (OCR) and the same examination session (May 2018). The seventh question was taken from a later session of the same paper from the same examination board. Comparison of the specifications for GCSE Mathematics of the three largest examination boards in England – OCR, AQA and Edexcel – indicates that coverage of topics is similar ('content uniform across the three big boards,' Barton, 2014,

online). There are some differences in their approach to assessment: AQA uses some multiple-choice items, whereas OCR, according to Craig Barton’s summary, ‘have focused on more wordy questions. So, less abstract, more real-world. No Multiple Choice. Papers ramped in difficulty’ (Barton, 2014, online).

Questions 1 to 6, common to both the pilot and main studies, were taken from OCR GCSE mathematics papers 1 and 4, first set on Thursday 24 May 2018. Question 7, added in the main study to give a comparison question on probability, was from OCR GCSE mathematics paper 1, set on Tuesday 3 November 2020. Table 4 shows the location of the questions within their original papers. Foundation Tier questions 1, 2, 5 and 6 were from OCR (2018a); Foundation Tier question 7 was from OCR (2020a); Higher Tier questions were from OCR (2018b).

In GCSE mathematics, students can gain grades between 1 (lowest) and 9 (highest). There are differentiated tiers of entry: foundation and higher. Figure 20 illustrates the grades obtainable in the different tiers, and the overlap.

Table 4 - Location of GCSE Mathematics Questions

Survey Q number	Level	Month/Year	Paper ref.	Page	Question number
1	Foundation	05/2018	J560/01	7	10
2	Foundation	05/2018	J560/01	12	16
3	Higher	05/2018	J560/04	11	10
4	Higher	05/2018	J560/04	12	12
5	Foundation	05/2018	J560/01	5	7
6	Foundation	05/2018	J560/01	8	12
7	Foundation	11/2020	J560/01	5	6

Source: Author’s own.

Figure 20 - Representation of Grades in Mathematics GCSE Foundation and Higher Tiers



Source: Ogden (2015) online.⁶²

Five questions (four for the pilot study) were taken from the Foundation Tier paper (paper 1), and two from the Higher Tier paper (paper 4). In general, on GCSE mathematics papers, the less demanding questions appear earlier in the paper; demand and difficulty generally increase through the paper (Barton, 2014; He *et al.*, 2015).

Some questions were common to both papers. None of these overlapping questions was selected for this study. This was decided upon to make sure that a range of difficulty of questions was selected, and so that the Higher Tier questions selected would be most likely to be more difficult than the Foundation Tier questions. The order of questions in the main study survey was also deliberately different from the order in which they appeared on the original question paper. This was done so that students would have to consider each question separately, and not simply assume that the earlier questions would be less difficult than the later questions.

Eight questions were initially selected for the pilot study. All questions chosen carried between 4 and 6 marks, so that statistical methods could be applied to compare students' answers. Given the study's focus, the chosen questions were on those topics that are most likely to discriminate between the different students. Wroe, a mathematics teacher and blog author

⁶² <https://www.ocr.org.uk/blog/new-gcse-9-1-mathematics-tiering-and-content-shifts>, accessed 12.06.2022

for OCR, suggests that question topics most likely to discriminate well between students of different abilities will be:

- ‘algebra and topics underpinned by algebra:
 - direct and inverse proportion
 - growth and decay
 - algebra
 - graphs of equations and functions
 - mensuration that involve formulae
- questions that assess reasoning and problem-solving’ (Wroe, 2021, online).

Apart from graphs of equations and functions, these topics were all represented in the questions selected. To allow comparisons to be made between the questions, the questions selected were all posed in verbal form: no questions that used graphs or diagrams were selected. This removed one possible source of variation and discussion, namely, the differing skills of students to interpret information presented in diagrammatic form, compared with their ability to understand information presented in verbal form.

The presentation of questions in verbal form only – and the elimination of graphs as a topic – was also a pragmatic choice, as the mode of delivery for the pilot study questionnaire had to shift from paper to online for the pilot study as a consequence of the COVID-19 pandemic. Because the questionnaire was distributed and completed via Microsoft Forms, it was not possible to include diagrams in this medium, and there was no mechanism for students to submit graphical or diagrammatic answers. Questions chosen had, therefore, to be ones that were posed – and that could be answered – wholly in words and numbers.

To further ensure that students would be able to access questions in the study, thereby improving the construct validity of inferences made from their answers (in accordance with the views of Kane, 2012), a discussion was then held upon the eight initially chosen questions with experienced teachers of mathematics in the researcher’s own school, including the subject leader and a former subject leader. They suggested the removal of two of the questions. In one case, the question topic was an advanced one that few students had studied,

whilst with regard to the other it was felt that the question was one of the most demanding on the examination paper and that it would, as a consequence, be likely to be found to be very difficult by almost all students. This, it was felt, could lead to it being insufficiently discriminating as a survey item. Informed by the observations of peer-colleagues, these two questions were removed from the survey, leaving six to be considered by the students in the pilot study. An additional foundation level question, on probability, was added to the main study so that it would be possible to ask students to compare two questions on the same topic. The results of their deliberations are reported in Chapters 5 and 6, and further discussed in Chapter 7.

The response rates for the pilot and main study are shown in Table 5.

Students' estimated the difficulty of examination questions on a Likert-type rating scale, the categories of which are discrete (e.g. "very easy", "easy" etc.). Likert scales are ordinal scales, producing categorical, non-parametric data. Rating scales such as this are widely used in research and, in the view of Cohen, *et al.*, 'rightly so, for they combine the opportunity for a flexible response with the ability to determine frequencies, correlations and other forms of quantitative analysis' (2011, p. 387).

Table 5 - Response Rates for Preliminary and Main Studies

Study	Date	Responses	Population surveyed	Response rate
Pilot study	February –March 2021	192	1125*	17%
Main study	September 2022	97	97**	100%

*5 schools, each with year groups of between 200 and 250; estimate 225 students per school.
5 x 225 = 1125 students

** All the students in four classes who attended on the days of the survey were included. There were 240 students in the year group, so this represents 40% of the year group.

Source: Author's survey data

According to Schwartz *et al.* (1991), Krosnick and Presser (2010) and Champagne (2014) (all cited in Cohen, *et al.*, 2018, p. 483), rating scales that have a verbal label for each point are more reliable than rating scales which provide labels only for the end-points of the numerical scales, and they are preferred by respondents. I therefore used a 5-point Likert-type scale with verbal labels for each point. The five points were described as “very easy”, “easy”, “neither easy nor difficult”, “difficult” and “very difficult.” There is an assumption of unidimensionality implicit in the use of Likert scales, implying a single construct: in this case, the construct is the student’s perceived difficulty of the question.

Numbers were assigned to the labels to facilitate reporting and analysis (very easy = 1, easy = 2, etc.) but there has been no assumption of equal intervals: “difficult” is not twice as hard as “easy”, for example. Nor can it be assumed that every student applied the same meaning to “easy” or “difficult”: these terms are relative and impressionistic (see Cohen, *et al.*, 2018, pp. 480-2). Because there is not an assumption of equal intervals between points on the scale, the range of statistical analytical techniques that can be applied to the numbers derived from the questionnaires is limited. Where mean values are calculated, these indicate the approximate location within the scale; mode values are also reported.

It might reasonably be expected that students who rated a question “very easy” or “easy” should also score well on their answer to the question, and *vice versa*. In this way, the scale for estimates of question difficulty and the scale for marks gained work in opposite directions. If correlations are to be observed between students’ estimates of difficulty and the marks they gain for their answers, it would therefore be expected that these correlations would be expressed in negative terms.

4.5 Study samples

My intention as a researcher was to recruit a sample of a reasonable size for the pilot study, and to base it across more than one school, in order to see whether there were differences in response between different school populations. I did not know what the response was likely to be, from students in other schools who did not know me, so I approached the headteachers of 14 schools in North East England to take part in the pilot study. My plan had been to visit participating schools in person, to explain the project and answer questions. However, the developing situation (globally and nationally) with regard to the COVID-19 pandemic ultimately meant that the survey was collected whilst schools were subject to the second national COVID-19 lockdown (January to March 2021). As a consequence of the well documented changes to both teaching and learning environments under the pandemic (DfE 2020), many schools closed their doors to all but essential visitors. Educational research, however, interesting and potentially impactful, was not included in the restricted list of activities allowed by most schools in the North East of England, who were by then focused on trying to make up for learning time lost during lockdown. The schools that ultimately provided the basis of the pilot study were those nearest to my own school, where I could use personal contacts to help facilitate agreement to allow the data collection (in a revised online format) to go ahead.⁶³ From 14 schools, 8 headteachers responded positively and, thereafter, curriculum leaders from 5 of the 8 schools agreed that their students could take part in the main study. Online questionnaires were distributed and, in total, 192 student responses were received. Once responses where students had not given their consent were removed, a total of 131 valid responses remained. Timelines for recruitment and data collection for the pilot study are collated in Table 6, with the different steps identified and described.

⁶³ The nature of the data collection method had to change in the face of unprecedented circumstances, following the strategic and adaptive approach outlined by Ashley (2021) and the 'pragmatic roots of mixed methods research' (Cohen *et al.*, 2018, 38)

Table 6 - Pilot Study: Timeline of Recruitment

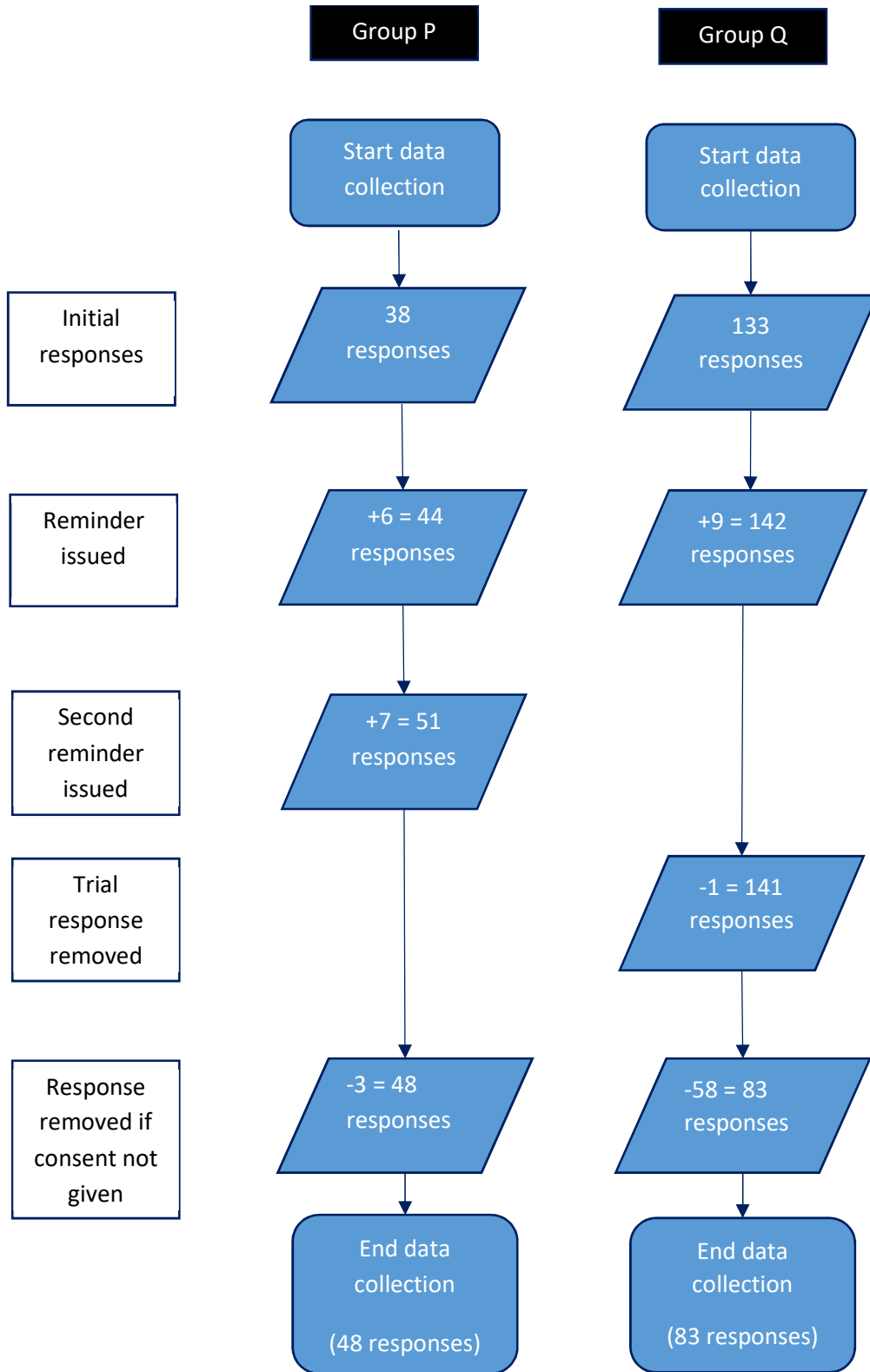
Step	Date	Action	Notes
1	28/02/2021	14 Headteachers contacted via email with personalised letter	12 comprehensive schools, 1 Pupil Referral Unit, 1 Special School, in Gateshead and North Tyneside Local Authorities
2	28/02 – 12/03/2021	Replies received from 8 headteachers	All replies indicated willingness to take part. All schools responding are mainstream comprehensive schools in Gateshead Local Authority
3	16/03/2021	Emails sent to Mathematics Curriculum Leaders of the 8 schools whose headteachers replied (step 2), with link to student survey via MS Forms	
4	17 – 18/03/2021	38 responses received from students at School A	School A is in Group P (these gave estimates of difficulty first, then answers to maths questions)
5	18 – 21/03/2021	Individual online meetings, phone calls and email exchanges with 3 curriculum leaders in Schools C, D, and E, to discuss the study	
6	24/03/2021	Reminder email sent to curriculum leaders of Schools A – E	
7	24 – 25/03/2021	6 further responses received from students at School A	
8	28/03 – 06/04/2021	133 responses received from students at Schools C, D, and E	Schools C, D and E are in Group Q (these gave answers to maths questions first, then estimates of difficulty)
9	21/04/2021	Reminder email sent to Schools B – E	
10	21 – 28/04/2021	9 further responses received from students at Schools C, D, and E	
11	30/04 – 13/05/2021	7 responses received from students at School B	School B is in Group P

Source: Author's own

In the pilot study, students were divided into two groups.⁶⁴ Group P students were asked to make Predictions as to the difficulty of questions first (and to give reasons for their predictions) before then attempting to answer the questions. In contrast, Group Q students were asked to answer the Questions first, and then estimate their difficulty and to give reasons. It was important for students to answer the questions as well as estimate their difficulty, so that a sense of whether they could in fact answer the question might subsequently be gained. The differences between the perceived difficulty and demands of a question and a student's performance in answering the same question are important. A student might, for example, describe a question as 'easy' but be mistaken and not then be able to answer it correctly. Another student, though, might label a question 'easy' because they could answer it without difficulty. These differences in self-knowledge are crucial for students and their teachers to understand, and they will form an important part of the learning models that will be discussed in Chapter 7. Figure 21 illustrates diagrammatically the timelines for recruitment and data collection for the two groups in the pilot study.

⁶⁴ Students from two schools were allocated to Group P, and students from the other three schools were allocated to Group Q. It was considered impractical to allocate students randomly to the two groups – and in any case the researcher had no foreknowledge of which students would choose to respond to the survey – so instead schools were randomly allocated to groups. This reversing of the order of answering in the two groups was decided upon because it was not evident beforehand whether students would be able to estimate the difficulty of a question without first attempting it.

Figure 21 - Data Collection Flowchart for Pilot Study Groups



Source: Author's own.

For the main study, recruitment of students was much more straightforward. Having struggled with adverse circumstances in the pilot study, I decided, on the advice of my supervisors, to focus on students from the school where I am the headteacher. If there were to be a bias towards participation, it would apply equally to all students recruited. Students in my school are streamed but not setted for mathematics, meaning that students of a range of broadly similar prior attainment are taught together in a class. Since my study survey included some questions from the higher tier mathematics papers, some of the topics of which are not taught to students who will take the foundation tier papers, I consulted with the head of the mathematics department at my school, and approached the class teachers of four classes in the upper and second streams, all of whom were predicted to gain grades between 5 and 9 in the mathematics GCSE in which they would be examined in the following May. All four teachers were happy for their classes to be approached. I visited the classes during their normal scheduled lesson time, and explained the study to the students. Students were given the opportunity to ask any questions, and they were given the opportunity to opt out of the study. All the students chose to participate. Owing to illness, some students were absent from school on the days of the survey, so the numbers in each class vary. Class numbers and survey dates are shown in Table 7, below. Following the pilot study, where students had been assigned anonymous codes beginning with the letters P and Q, students in the main study were assigned anonymous codes beginning with the letter R.

Table 7 - Data Collection for Main Study

Class	Survey Date	Responses	Number of students
11Y/2	21.09.2022	R72-R95	24
11Y/1	23.09.2022	R01-R27	27
11X/1	27.09.2022	R28-R53	26
11X/2	27.09.2022	R54-R71	18
			Total: 95

Source: Author's own

Table 8 - Samples and Gender Distribution

Pilot Study								
Group	School	Male		Female		Non-binary / prefer not to say		Total
		No.	%	No.	%	No.	%	No.
P	A	14	23.7%	25	38.5%	0	0.0%	39
	B	3	5.1%	1	1.5%	1	33.3%	5
Q	C	8	13.6%	10	15.4%	0	0.0%	18
	D	26	44.1%	13	20.0%	0	0.0%	39
	E	8	13.6%	16	24.6%	2	66.7%	26
Total		59	46.5%	65	51.2%	3	2.3%	127

Main Study								
		Male		Female		Non-binary / prefer not to say		Total
		No.	%	No.	%	No.	%	No.
Total		41	43.2%	53	55.8%	1	1.1%	95

Source: Author's own.

The gender distribution for both studies is given in Table 8 (above). Overall, the distribution was roughly equal between male and female students.

4.6 Focus Groups

In order to find out more about the experiences and comprehension of students in relation to demand and difficulty in examination questions, I chose to hold discussions with small groups of students from the survey groups, to gain their individual and collective views. As Gibbs notes, the distinguishing feature of focus groups – in contrast to individual interviews – is ‘that they are *interactive*, the group opinion is at least as important as the individual opinion, and the group itself may take on a life of its own not anticipated or initiated by the researcher’ (Gibbs 2021, p. 240, emphasis in original). So it was in this study: the group dynamic enabled particular lines of enquiry and student interest to be followed and elaborated upon, which in turn led within the reflexive thematic analytic stage to the development of a particular theme (“motivation”) that had not been initially developed from questionnaire responses. As described by Halcomb *et al.* (2007) and Shelton *et al.* (2019), the focus groups in this study were used for a number of purposes, including the co-construction of new knowledge, gauging

student opinion, hearing the voices in research of students who are not normally represented, and learning from students' experiences. On a practical note, it was straightforward for me, as headteacher-researcher, to recruit volunteers to a focus group and to conduct the interviews, since this could be done during school time. The students involved in the groups appeared to benefit as well (Gibbs, 2021, p. 241): as can be seen from the transcript and quotations, they clearly enjoyed the group dynamic, they were able to talk about aspects of learning in ways that they would not do otherwise, and the interviews gave them the opportunity to debate their approaches to examination questions ahead of their 'mock' and formal examinations. Halcomb *et al.* (2007) and Barbour (2018) observe that the researcher, as focus group facilitator, plays a critical role in determining the experiences and outcomes of the group, and recommend that the researcher needs to be knowledgeable about group dynamics and skilled in group facilitation. As the students' headteacher, well used to speaking with groups of students both formally and informally, I was able to create an atmosphere with high levels of trust and respect, and in any case, examination questions in mathematics are a 'safe' topic for teenagers to discuss. Potential problems concerned with confidentiality (Sim and Waterfield, 2019) therefore did not arise. Because I was asking student volunteers in my school about matters connected with teaching and learning, and because 8 out of 10 students had already been involved in an earlier part of the study, organisational and consent issues were handled without difficulty. Students also gave their verbal assent at the start of the interview.

Following the marking of the mathematics questions and initial sifting of questionnaire responses, two focus groups of students were recruited, from two of the four classes. Each group comprised of five students, all volunteers (various authors, as collated by Gibbs, 2021, advise group sizes of between 4 and 12). Focus Group 1 was made up of five male students from class 11X/1, all of whom had previously completed the questions and survey. These were students R31M, R34M, R36M, R49M and R51M. This discussion took place on Monday 19 December 2022 at 11:30am. Focus Group 2 was made up of two female and three male students from class 11Y/2. Three of these students had previously completed the questions

and survey. These were students R73F, R77M and R87M. The remaining two students had been absent on the day of the survey, and they had not completed the questions or the survey. These two additional students were given the codes R96F and R97M. The discussion with Focus Group 2 took place on Tuesday 20 December 2022 at 10:30am.

The focus group interviews were conducted as semi-structured interviews. That is, there was a basic set of questions, which had previously been discussed between the researcher and supervisors, and these questions were asked in both groups. Follow-up questions were asked in each group, in response to statements made by students. These questions are reproduced below. Copies of the mathematical examination questions from the survey were available to the students in the focus groups, so they could relate the discussion points to the actual questions.

Focus Group Questions

1. When you look at an examination question in maths, what do you look at first? What is your focus? What happens next/after that? And then? Do you always use this same approach or would your approach be different in an examination from how you would approach a question in class?
2. Can you talk me through how you would solve question number 1? [This is a question that most people answered correctly]. Follow-up questions to seek clarification.
3. Can you talk me through how you would approach question number 3? [This is a question that many people did not answer correctly.] How did you develop your thinking for answering this question? Follow-up questions to seek further clarification.
4. [Present longer more 'wordy' question and a shorter question.] How would your approach differ for these different sorts of questions? Which type of question do you find easier to understand and solve? Can you explain this? Why might some other people disagree?
5. I am interested in how you think students (like you) improve your skill in mathematics. Can you talk to me about this? How did you develop your mathematical skills? What did you do to understand better when you had particular difficulties?
6. How do you think teachers can make maths more interesting and understandable for students?

Focus group interviews were recorded, with the students' consent. Focus Group 1's recording was 33 minutes long; Focus Group 2's was 21 minutes long. The recordings were then transcribed verbatim by the researcher, including hesitations, fillers and grammatical slips. Anonymous codes were substituted for students' names, including in places where the students referenced one another. The verbatim transcripts appear as Appendices E and F. Recordings will be retained by the researcher for one year from the end of the research project and then deleted. The transcripts of the focus group interviews were analysed using the same method of reflexive thematic analysis as the responses in the pilot study and main study questionnaires. As Barbour (2018) and Wilkinson (2011) suggest, it needs to be borne in mind during the analytical phase that the unit of analysis in the focus group is the collective perspective, not an individual conversation between the researcher and each individual participant. Quotations from the focus groups are therefore given at some length in Chapter 6, with the voice of the researcher and contributions from different group members also included, in order to properly represent and capture the group dialogue.

4.7 Data analysis techniques: descriptive and inferential statistics

As briefly alluded to in Section 4.1, this study is primarily qualitative in focus, in keeping within constructivist principles, and so it principally collected free text responses through a questionnaire, and students' spoken thoughts through focus group interviews. However, it also collected some quantitative data through students' answers to mathematical questions and through their responses to Likert-type scale questions as well as qualitative data through free text responses. Where the data collected was empirical in nature, this was analysed using standard statistical methods on Microsoft Excel spreadsheets, whereas where the data was qualitative, a reflexive thematic analysis approach was used.

In the analysis of students' responses to the questionnaires, descriptive statistics are given, in keeping with the approach advanced by Field (2009), for frequencies (numbers and percentages), and central tendencies (mean and mode averages) of the distributions of

answers. Measures of dispersal (standard deviation) are given. These descriptive statistics are reported and illustrated, using tables.

Correlations will be explored, statistically using the Pearson product-moment correlation coefficient (Field, 2009)⁶⁵, and graphically using bubble charts (Battista and Cheng, 2011); the possible inferences of any correlations found will be discussed.

As well as these descriptive statistics, inferential statistical techniques were employed so that properties of the population that might exist beyond the immediate data sets of the study could be inferred. Accordingly, in the main study, following the approach outlined by DeCoster (2006), a t-test was performed to test whether there was a statistically significant difference between the estimates of the two groups (P and Q). In addition, and so that the findings of the study might be applied to other real-life patterns of student behaviour in tackling examination questions, crosstabulations were performed to investigate whether there was a difference between the response patterns of male and female students across the whole study, and between the sexes in the two groups.

CRAS scales (see Section 3.2.4) are used to evaluate the demands of the GCSE Mathematics questions chosen for the study. This evaluation of demand provides a link from the literature to the discussion of demand and difficulty as experienced by students in Chapters 5 and 6.

Having applied basic statistical measures to the empirical results, the free text responses were then subjected to a process of thematic analysis so that underlying themes might be constructed, developed and explored.

⁶⁵ 'The Pearson product-moment correlation coefficient (Pearson correlation coefficient, or 'Pearson's r ', for short) is a commonly used measure of the linear relationship between two variables. In essence, it attempts to create a line of 'best fit' through the data points of the two variables, and the Pearson correlation coefficient shows how far away all these data points are, on average, from this line of best fit. The value of r can be positive, but it can also be negative, showing an inverse correlation between two variables' (Field, 2009, pp. 166-174).

4.8 Data analysis techniques: reflexive thematic analysis

To enable analysis of the free text responses of the students which made up the bulk of the data collected in all the parts of this study, a method of thematic analysis was used. This approach was chosen because it offers ‘an accessible and theoretically flexible approach to analysing qualitative data’ (Braun and Clarke, 2006, p. 77). More precisely, this present study follows a *reflexive* thematic analysis approach; Braun and Clarke (2022, p. 5) explain that reflexivity involves the practice of ‘critical reflection on your role as researcher, and your research practice and process.’ Braun and Clarke have been particularly influential in developing this technique, as researchers, writers and journal editors. Their 2006 journal article ‘Using thematic analysis in psychology,’ in which they explained and codified the method, has been cited over 170,000 times, according to Google Scholar⁶⁶. They and others have used reflexive thematic analysis especially in studies with more marginalised groups, including feminist, first nation, minority ethnic, and LGBTQ+ communities, to make present the views and stories of those who are less often heard and represented in social science research. As we have seen in our review of literature (section 3.4), the voices of secondary school students are rarely sought or heard in discourses around formal assessment or the evaluation of teaching and learning; it seems appropriate, therefore, to adopt Braun’s and Clarke’s reflexive thematic analysis technique in this study.

Since the publication of Braun’s and Clarke’s 2006 landmark article on thematic analysis, and particularly since their delineation of ‘reflexive thematic analysis’ in 2019, the technique of reflexive thematic analysis has been used in a handful of studies focused on the experience of children and adults in UK secondary schools. The subject matter for these studies has included inclusion and exclusion (Murphy, 2022; Martin-Denham 2021; Dunleavy and Sorte, 2022), the teaching of health education and mindfulness (Pickett *et al.*, 2017; McGeechan *et al.*, 2019),

⁶⁶ Google Scholar, accessed 21.08.2023. ‘Since initially writing on thematic analysis in 2006, the popularity of the method we outlined has exploded,’ Braun and Clarke, 2019, p. 589.

school travel (Nikitas *et al.*, 2019), and coaching and counselling in schools (Anthony and Van Nieuwerburgh, 2018; Lynass *et al.*, 2012). The impact of COVID-19 lockdowns on students and teachers in UK schools has generated the single largest cluster of studies, albeit still a modest number: De Carvalho and Skipper, 2019; Kim and Asbury, 2020; Kim *et al.*, 2021; and Taylor-Egbeyami *et al.*, 2023. All the studies listed here have cited Braun and Clarke (2006 and/or 2019) as influences on their research and analysis methods. It is possible to observe, therefore, that the technique of reflexive thematic analysis is becoming established as an effective way to tell the stories and experiences of school students, particularly in response to aspects of their vulnerability and mental health. However, literature searches using university library search facilities, and using Google Scholar and ERIC, revealed no published studies using reflexive thematic analysis that were related to the topics of teaching and learning or educational assessment. This study therefore breaks new ground, at the same time as it builds on and complements the work of authors with similar interests in presenting and interpreting the lived experiences of secondary school students.

Reflexive thematic analysis falls within the values of a fully qualitative research paradigm, named 'Big Q' by Kidder and Fine (1987). Braun and Clarke (2022) identify certain orientations and skills as characterising this approach, distinguishing it from a quantitative paradigm. Their analysis is summarised below (Table 9), with reflections added from the current study.

By contrast, much educational research is currently situated within a quantitative-positivist paradigm, where often a hypothesis is formed and rigorously tested using large-scale datasets. Randomised control trials (RCT), which are one of the most rigorous manifestations of this quantitative-positivist methodology, and which have become regarded as the 'gold standard' in medicine (Hariton and Locascio, 2018), have also become more widespread in education.⁶⁷

⁶⁷ For instance, Camilla Nevill, the Education Endowment Fund's Head of Evaluation, asserts that 'RCTs are currently the optimal and least-biased method for estimating, on average, whether something works, when done well'; that they 'provide powerful information for [educational] decision makers'; and that they are popular with schools: EEF-funded projects have recruited over 13,000 schools to RCT projects. From her contribution to Behavioural Exchange 2019 Conference (online), 31.10.2019

RCTs are particularly used for evaluating the effectiveness of an intervention. As educationalists have become better informed by the science of learning, more scientific methods of research, such as RCTs, have become highly regarded in educational circles. But while large-scale quantitative studies can provide useful generalisable information for leaders, they cannot (and do not seek to) tell the stories of the students and teachers who work within these systems. There is a place for immersive qualitative study, such as the present project, especially since very few educational researchers have the opportunity to be embedded in the culture they seek to study.

Table 9 - Key Differences between Qualitative and Quantitative Paradigms, and their Application to this Study

Aspect of research	Qualitative paradigm	Quantitative paradigm	Application to the current study
Research purpose	Broadly focused on <i>meaning</i> – understanding situated meaning. Aims to generate contextualised and situated knowledge	Recording and understanding truth; often seeking explanatory models or theories. Often reductive, often hypothesis testing	Focused on understanding situated meaning of student’s experiences; aiming to generate knowledge that is situated within the context of secondary school education
Big Theory positions related to how ontology and epistemology are understood	An only-ever partially knowable world, where meaning and interpretation are always situated practices. Non-positivist; multiple and varied theories (e.g. constructivist, critical realist)	A world knowable through systematic observation and experimentation. Positivist or postpositivist; realist.	Constructivist approach taken, interrogating and interpreting students’ reported views to co-construct situated understanding. Multiple theories are brought into the literature base for this study, illuminating different perspectives
Orientation to truth	Situated or life-embedded truth, partial truth, multiple truths	Singular truth	Situated and multiple truths
Researcher role	<i>Situated</i> interpreter of meaning; subjective storyteller. Subjectivity is valued	<i>Impartial</i> observer of object of study; unbiased reporter. Objectivity valued, which subjectivity threatens	Interpreter of meaning, embedded and situated within the school world as headteacher-researcher
Researcher subjectivity	Not just unproblematic, but an asset, especially if reflexively engaged with	Introduces bias which threatens analytic validity; requires measures to control	Subjectivity is inevitable, given the situated nature of the headteacher-researcher. This is regarded as an asset, although reflexively interrogated and balanced with perspectives from colleagues and supervisors
Orientation to influence of subjectivity	Reflexivity as a tool to both interrogate and harness the value of subjectivity	Bias control measures to reduce or eliminate influence	Reflexivity used as a tool to interrogate and get value from this dual role

Aspect of research	Qualitative paradigm	Quantitative paradigm	Application to the current study
Data purpose and sampling	To gain rich, in-depth understanding; smaller samples valued	Ideally to gain generalizable understanding; larger, representative samples ideal	To gain rich, in-depth understanding. Some larger samples used to contextualise and validate the study; smaller samples valued for more in-depth exploration within focus groups
Data analysis	Focused on text and meaning	Focused on numbers; relationships between variables, cause and effect	Some basic statistical analysis of qualitative data, to provide context and overview. Major part of study focused on text and meaning
Contributions to knowledge	Part of a rich tapestry of understanding	Stepping-stone towards complete or perfect understanding	Seeking to contribute to a rich understanding of student's experiences of formal assessments, within the knowledge domain of teaching and learning

Source: First three columns – summarised from Braun and Clarke (2022, p. 6); right-hand column – Author's own.

Elucidating further on the practicalities of such an approach, Braun and Clarke outline six steps to thematic analysis that seek to ensure deliberate and rigorous analysis. This six-step approach was applied to both the pilot and main studies undertaken for this thesis, as detailed in Table 10.

Table 10 - Thematic Analysis in the present Study

Phase	Notes relating to the present study
1. Familiarizing yourself with your data	Downloading survey data (pilot study), sorting and printing; reading students' responses. Reading students' written responses on paper (main study); Re-reading to ensure familiarity with range and texture of the data. Making notes of initial ideas
2. Generating initial codes	Generating an initial list of ideas about what is in the data and what is interesting, noting examples of coding from extracted data. Coding examples appear below
3. Searching for themes	Sorting different codes into potential themes, using visual representations. Some codes may not appear to belong anywhere; nothing discarded at this stage
4. Reviewing themes	Sifting: some potential themes may not be strong enough to become themes; there may not be enough data to support them, or they may be very close to other themes. Other themes may be dense and need splitting. Data should cohere within a theme, and there should be clear and identifiable distinctions from other themes. Validity considerations: do the identified themes represent the meanings across the whole data set? Reviewing as an iterative process that will be stopped when nothing new is emerging. An example of reviewing a theme is given in Chapter 5
5. Defining and naming themes	Defining and naming themes. A satisfactory thematic map should have been developed by this point. Define and refine to identify 'essence' of each theme. For each theme, conduct analysis and write a description identifying the theme's 'story', explaining how it fits into the project, how it engages with research in the literature review, and how it creates arguments that relate to research questions. Themes are named and defined in Chapters 5 and 6
6. Producing the report	Aiming for a concise, coherent, logical and interesting account that tells the complicated story of the data. Need to convince the reader of the merit and validity of the inferences made. Choosing vivid examples that capture the essence of each theme; embedding sufficient of these data extracts within a compelling analytic narrative. Analysis will inform the exposition of an argument in relation to the research question: Chapter 7

Source: (left-hand column) Six steps from Braun and Clarke (2006); (right-hand column) Author's own explanatory notes.

In both parts of the present study, I adopted each of these steps. At the same time, I kept a reflexive journal, which I reviewed regularly. This was valuable in allowing me to see how my perceptions and understanding of the data grew and changed over the months of analysis. Qualitative data were reviewed and coded. Coding, Braun and Clarke stress (2022, p. 54), is an 'organic and evolving process... Coding begins without any list or set idea of what codes will be used.' My first step was to read through the questionnaires, repeatedly. Coding is about spotting the connections, the repetitions of meaning in what students have independently written. I started to note down on large pieces of paper phrases that appeared to cohere together. Having read through my main study data set a number of times, I began to see that many students were commenting on the *wording* of questions and how that related to how difficult they found them, and that the *methods and different steps necessary* to address the question also featured frequently. I started to highlight these different candidate themes in different colours on photocopies copies of the student questionnaires. An example appears below, Figure 22. In this example, a yellow highlighter has been used for comments about methods and different steps, and the ways in which they contribute to the difficulty of a question. A pink highlighter has been used to indicate comments relating to the wording of questions.

Figure 22 - Example of Main Study Student Questionnaire, with Highlights

b) Which question did you rate as the least difficult? Q5

Please give your reasons

the question was laid out ~~easy~~ simply and it was easy to comprehend, what I needed to work out for the answer.

c) Which question did you rate as the most difficult? Q3

Please give your reasons

don't know what method would be used to work it out, I also had a harder time figuring out what steps would have been required to find the answer.

d) Questions 2 and 7 are both about probability. Which one do you think is more difficult? Q7

Please explain why you think this is

I found the question itself to be easier, however it took longer to realise the simplicity of the question due to having over complicated it at first and looking for methods that would have been unnecessary.

e) What factors do you think make examination questions in Mathematics more difficult?

complicated wording of questions, added ~~was~~ unnecessary information that makes it appear harder than it is.

Source: author's own

Although 'a single coder is normal – and good – practice in reflexive TA' (Braun and Clarke, 2022, p. 55), I also shared and checked examples of coding in both studies with the head of the mathematics department at my school, to gain validation through a second opinion, in line with an approach suggested by Braun and Clarke (2006).

To give an insight into my analytical process of familiarisation, coding and refining that led to the development of this theme, I have included an extract of one of my coding sheets as Figure 23, below. On this sheet I have collected together all the student comments from questionnaires that appear to relate to aspects of memory, practice and familiarity. The situated nature of my research, and the deeper levels of discussion that ensued through the

focus groups, allowed me to take an organic, evolving and subjective approach to the coding process, as outlined by Braun and Clarke. In the case of this theme, the coding was largely inductive: I was working with the students' comments, reflecting on and wanting to find out more about their experiences and perspectives, and interpreting them to construct a theme. However, the 'Big Q' approach, taking account of the researcher's subjectivity, means that this cannot be a 'pure' inductive orientation, because of what I 'bring to the data analytic process, as [a] theoretically embedded and socially positioned researcher' (Braun and Clarke, 2022, p 56). This theme is grounded in cognitive load theory and the working memory model; continual reference to these theoretical frameworks was an essential part of my reflections during coding.

Figure 23 shows an early stage in the analytical process, where comments with common content were gathered together. Coding is exploratory at this stage – all potentially relevant ideas are being kept in play, because the eventual direction of the theme is not yet known. The extract also shows a further level of coding and analysis, that of initial theme development, where some of the complexities of the comments are starting to be separated out visually, using different coloured highlighters. There is no standard way of doing this, but this manual and visual method worked well for me. I worked through the dataset systematically several times, spotting and labelling new things until I felt that the different meanings had been well captured and differentiated. I then transferred the colour-coded comments to a fresh sheet and continued with the provisional development of themes.

Figure 23 - Extract from Coding Sheet for "Memory, Practice, Familiarity"

memory
memorise
remember

forget/forget
16

25
10
3

Memory, Practice, Familiarity

Coding

forgot; hard to memorise

practise

memory, memory test

Student	Comment	Question
05F	[exams may not distinguish correctly between students of differing abilities, because] less able students could be given exam questions which they haven't revised for, or haven't learned how to do [infers these would be too hard, or wouldn't show their ability]	
06M	Forgot how	5
090M, 10F, 16M	Unsure about the equation	6
10F	[least difficult] because I found that I remembered it when it was taught	
13F	[questions challenging] especially when worded differently to how I revised it	
14F	I couldn't remember how to do it that well Can't remember/don't understand how to work out the circumference	2 6
15N	[difficulty affected by] whether or not I revised [Not difficult, just] I didn't remember how to find the circumference [probability tree] I kinda forgot how to do one because it had been so long	6 2
17F	[easy] because we've learned it at the start of secondary and have constantly had practice by either revising it in lesson or using it to solve other questions	Q5+7
19F	We have been learning how to probability and factorising for a long time. It is just easy	Q1,2,5,7
19F	Hard to memorise the equations for circles and general equations in maths	Q6
20F	Basic maths which I've know how to do for a long time	Q1,4,7
22F	[Q2 difficult because I] couldn't remember how to complete the question	2
22F	There is also a lot that needs to be remembered which can get difficult and confusing	
24M	[Q2 difficult] Because the diagram needs to be involved for Q2, which I can slip up on easily due to the amount of things you need to apply to it and what you need to remember about it	
26F	[easy because] it just retrieval of given information [factors that make Qs difficult] forced equation memorisation Sometimes people just don't remember stuff as well as other people, then its just a memory test, not a knowledge test	7
27M	I think if you revise a lot then you can pick up questions a lot easier	
29F	I personally, find it alot of pressure in exams and forget alot.	
33F	[Q2 is more difficult than Q7 because] I forgot how to use probability trees and I didn't like how it was worded	
35F	[Q5 easiest] We cover this topic a lot in maths. Also this is a required skill in other topics so you need to know how to do it well to get higher stakes questions right. So there is more practice on this.	
36M	[hard] when knowledge of content needs to be put into a context that has not explicitly been practised in lessons	
37M	[Qs 1,2,4,5 and 6 less difficult because] I've already practiced them a lot	1,2,4,5,6

worded

worded

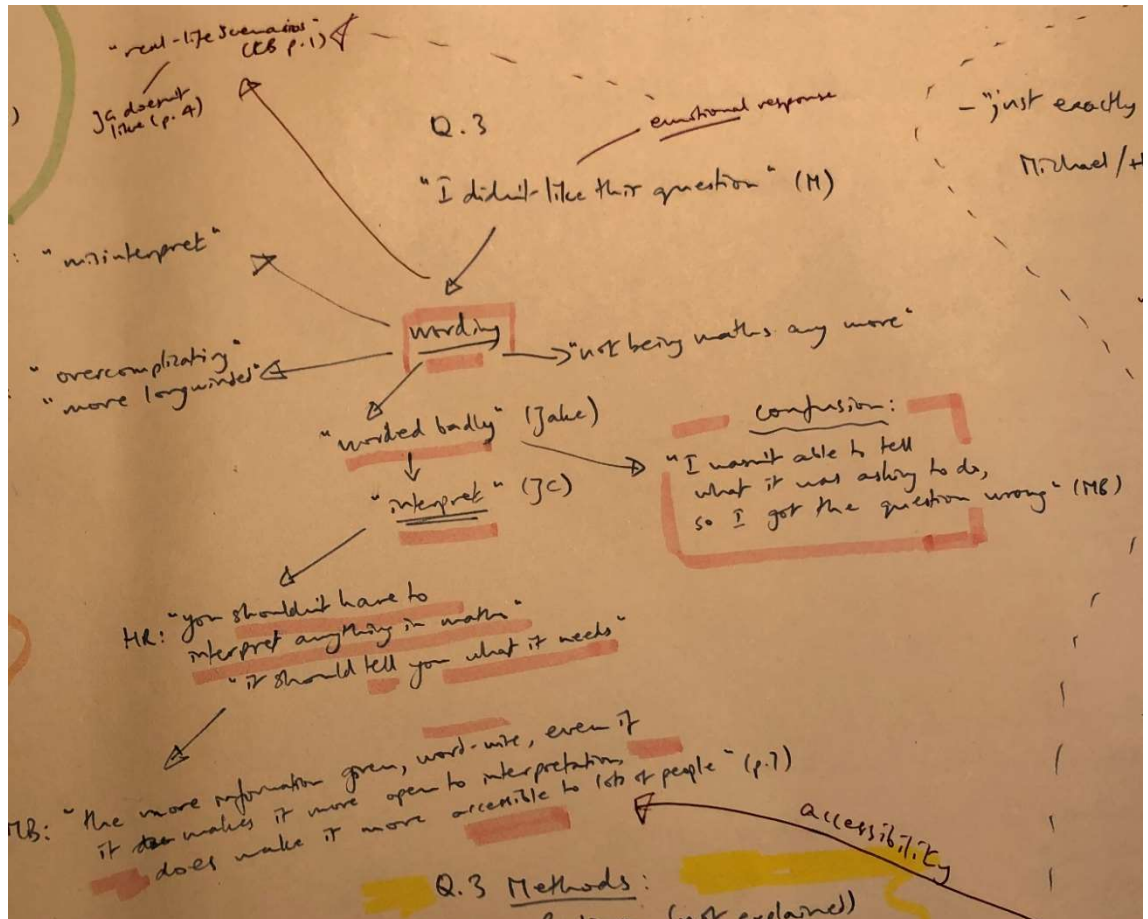
worded

Source: author's own

Following this stage in the process, I created some visual thematic maps. Large pieces of paper were used to arrange and rearrange comments and questionnaire fragments as I searched for themes. An example of this visual representation is included as Figure 24 (below). This example comes from the focus group discussions, where students' comments around wording and confusion were starting to be developed into a possible theme. This is a photograph of a

small area from a much larger sheet of paper, where the boundaries between possible themes are still fluid and developing. The edges of other possible themes can be seen (green and yellow pen) on the margins.

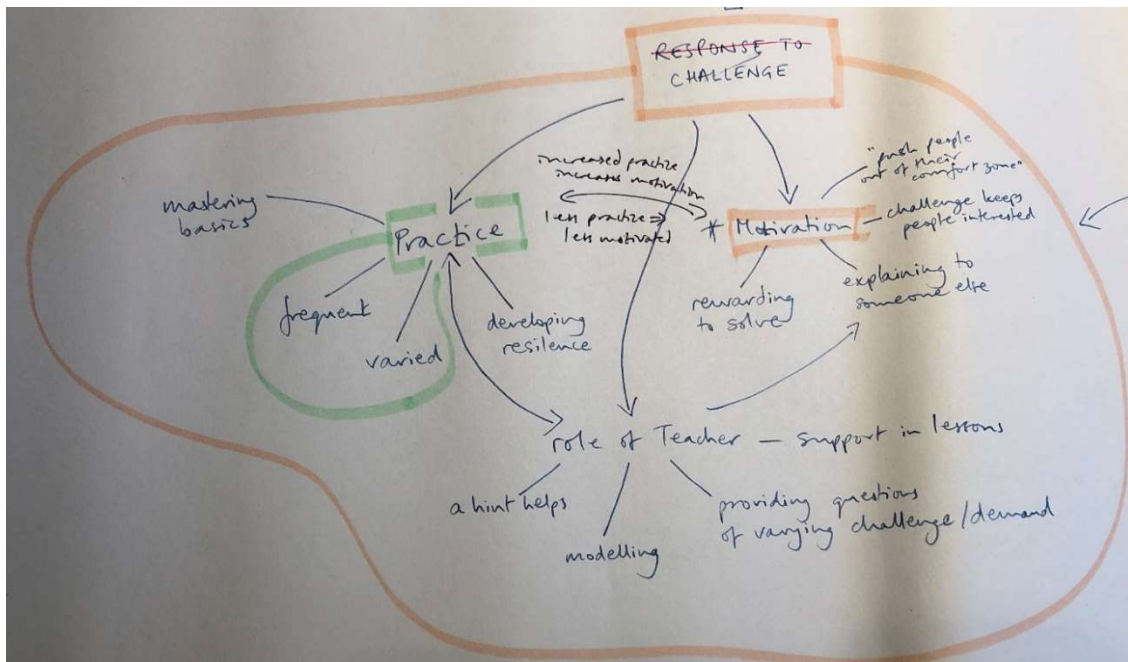
Figure 24 - Example of Visual Arrangement of Student Comments from Focus Groups



Source: author's own.

Once I had started to develop themes, I used visual representations to collate and review them. These developing themes were discussed with my principal supervisor, who has experience in qualitative analysis. An example of this visual representation appears as Figure 25, below.

Figure 25 - Mapping Developing Themes



Source: author's own.

As can be seen from Figure 25 above, I was not sure at this stage whether the possible themes of “practice” and “motivation” were linked, and whether they were both sub-themes of something that might be called “responses to challenge.” The role of the teacher is also under consideration in this example of thematic mapping. These three examples show clearly that my role as researcher was quite inductive and data-led, although through the reflexive qualitative process I brought to the meaning-making my experience of examinations as a teacher and examiner, my understanding of learning theory, and my familiarity with the students; this gives my analysis a thorough grounding in the world of the student.

In reflexive thematic analysis, coding ranges from semantic to latent; these can be thought of not as opposites but as two ends of a spectrum (Braun and Clarke, 2022, p. 57). Much of the coding in this current study lies at the semantic level, being participant-driven and descriptive, exploring meaning at the surface level of the data. It is important to stress, however, that, as the researcher, I had a very *active* role in generating these themes. Braun and Clarke (2023, p. 3) remind us that, ‘in reflexive TA [thematic analysis], themes are generated, created or

constructed (for example), they are not identified, found or discovered, and they definitely don't just "emerge" from data.' There was one set of codes and themes – that eventually became the "motivation" theme – that were developed at a more latent level, because their meaning seemed to lie at a deeper, more implicit and conceptual level. As I reflected more on the interplay between students' comments, particularly in the focus groups, I became aware of the links between what the students were saying and my reading of theoretical and professional literature (particularly Jackson, 2010, on fear in education, and McCrea, 2020, on motivated teaching); this developing theme became more researcher-driven and conceptual. Over time, and as a result of my discussions with my principal supervisor, these themes were refined and developed, mapped and named. These themes were continually related back to the research question. The construction of thematic maps led to further refinement, including re-naming and further analysis. The final stage was to write the story of each theme, and to bring a rich exposition, illustrated with excerpts from students' written and spoken responses, into the Discussion in Chapter 7.

4.9 Summary of the philosophical approach and methods in this study

This study takes an epistemological stance of what might be termed 'tempered realism' (Smith *et al.*, 2017, p. 564). That is, there is assumed to be a broadly uncomplicated relationship between the language that students have used in their responses and the reality that they are meaning to convey. Most of their responses have been analysed at the semantic level, therefore. At the same time, as both headteacher and researcher, I acknowledge that the interactions between researcher and participants will impact on both parties. In this way, my values and assumptions contribute (to a greater or lesser extent) to the formation of the questions I ask, the ways in which students respond, and the way that I read the data. For all these reasons, a reflexive thematic analysis approach (as exemplified by Braun and Clarke, 2022, see section 4.8 above) felt most appropriate for this study. By adopting such a reflexive

thematic analytical approach, I have acknowledged the importance of continual reflection on my own subjectivity and the impact it has had on data collection and interpretation. The data was analysed using a “bottom-up” inductive approach, in which, although students’ responses were constantly reflected on against a theoretical framework of relevant published educational research, there was no attempt to force the students’ responses to fit within any pre-existing theoretical approach or to test them against a pre-determined hypothesis.

Having now set out in detail the methods that have been used in this study, the emphasis now turns to the new material reported, and to its interpretations. The next two chapters therefore present and interpret findings from the pilot study and the main study.

Chapter 5: Pilot Study – Embarking on my first proper trip

In this chapter I make my first “journey to the foreign land,” in the form of my pilot study. Owing to some “local difficulties” with disease and travel restrictions, I “find, then lose, then find my way again.” I reflect on the lessons I have learned from this initial trip.

The pilot study involved students from 5 comprehensive secondary schools in the North East of England rating the difficulty of, and attempting solutions to, 6 GCSE mathematics past paper questions. The survey was carried out by the researcher via the internet, using Microsoft Forms. As discussed in Chapter 4, students rated the difficulty of questions using a 5-point Likert-type scale, and they were invited to explain their choice of difficulty grade, using a free text response field. Students also entered their answers to the 6 mathematics questions.

Before conducting the pilot study, I did not know whether it would make a difference to students if they made their estimates of difficulty *before* or *after* attempting the GCSE mathematics questions. The pilot study was therefore divided into two groups, as detailed in Chapter 4. In group P, students made their Predictions of difficulty first, before giving their answers; in group Q, students answered the Questions first, before giving their estimates of difficulty. For the marks obtained in answering the examination questions, a *t*-test established that there was no significance between the mean scores of students in groups P and Q, $t(60) = -0.92, p = .361$. Groups P and Q were therefore amalgamated for analysis and discussion of their answers to the examination questions, with any variations noted between the two groups. Some small differences were observed between the groups, however, in terms of the patterns of the students’ estimations of difficulty, and these are reported in the discussion below (section 5.2).

5.1 Data analysis: descriptive statistics

Descriptive statistics are reported for the numbers of students and their distribution, by school and by gender, with tables showing distribution of student responses for estimates of difficulty

and for the marks they gained for their answers to the mathematical questions. From these, the spread and distribution of student responses are explored, correlations between the data sought, and inferences of the findings discussed. 127 students completed the survey, 44 from group P and 83 from group Q; the gender split was roughly equal. Average completion time for the survey questionnaire was 18 minutes, and there was a wide spread of time taken, from some very short times from the 65 students who did not consent or proceed to the pilot survey, to one student who took just over one hour to complete write his particularly full responses. Table 11 (below) shows the collated results of the students' estimates of difficulty. The topic for each question is included as well as a summary of the distribution of estimates and calculations of the mean and standard deviation relating to the distribution.

Table 11 - Pilot Study: Students' Estimates of Difficulty

Students' estimates	Q1 Algebraic expressions	Q2 Probability	Q3 Conditional probability	Q4 Growth and decay	Q5 Algebraic expressions	Q6 Mensuration	Total %
1 Very easy	20	16	6	3	25	37	14.0%
2 Easy	43	36	22	15	37	39	25.2%
3 Neither easy nor difficult	37	35	38	33	41	25	27.4%
4 Difficult	21	32	42	46	16	16	22.7%
5 Very difficult	6	8	19	30	8	10	10.6%
Total	127	127	127	127	127	127	
Mean	2.6	2.8	3.4	3.7	2.6	2.4	
Mode	2	2	4	4	3	2	
Standard deviation	1.1	1.1	1.1	1.0	1.1	1.2	
Rank by mean (most difficult first)	4	3	2	1	5	6	

Source: Author's own.

In Table 11, the range of difficulty estimated is also given, and the number of students making each estimate for each question is reported. It can be seen that “easy” was the most numerous estimate for questions 1, 2 and 6 (the numbers of responses have been emboldened), but “difficult” was the most numerous estimate for questions 3 and 4, showing that students found questions 3 and 4 more difficult than the other questions. A rank order of difficulty, sorting by the mean values, has been included with question 4 as the most difficult and question 6 the least difficult. Overall, the middle option, “neither easy nor difficult” was the estimate that was selected the largest number of times, taking 27.4% of the total estimates; this indicates a ‘central tendency’ and an avoidance of extremes (Cohen, *et al.*, 2018, pp. 483-4). This shows that students in the pilot study tended to avoid the extremes of the Likert-type scale. It is evident, however, that students’ estimates tended to be placed more towards one side of the scale or the other for each question (apart from question 2). Marks that the students gained in their actual answers to the mathematical questions, and their associated statistical measures, are reported in Table 12, below. The number of marks available per question varied: most questions had 4 marks available, but question 3 was out of 5 marks, and question 5 was out of 6 marks. As well as the marks that students gained, the proportion who attempted each question is reported, along with the proportion of those who attempted each question, and those who gained either no marks or full marks.

Table 12 - Pilot Study: Students' Marks for GCSE Mathematics Questions

Students' Marks and Attempts	Q1	Q2	Q3	Q4	Q5	Q6	
	Algebraic expressions	Probability	Conditional probability	Growth and decay	Algebraic expressions	Mensuration	
Marks:	0	22	78	52	40	4	39
	1	10	5	8	8	22	3
	2	12	2	0	7	4	2
	3	2	0	13	3	35	51
	4	61	13	0	10	3	
	5			14		9	
	6					20	
Attempts	107	98	87	68	97	94	
% attempted	84.3%	77.2%	68.5%	53.5%	76.4%	74.0%	
Fully correct	61	13	14	10	20	51	
% Attempts fully correct	57.0%	13.3%	16.1%	14.7%	20.6%	54.3%	
% Attempts gaining zero marks	20.6%	79.6%	59.8%	58.8%	4.1%	41.5%	
Mean mark	2.7	0.6	1.3	1.0	3.0	1.7	
Mode mark	4	0	0	0	3	3	
Standard deviation	1.7	1.4	1.9	1.5	1.8	1.5	

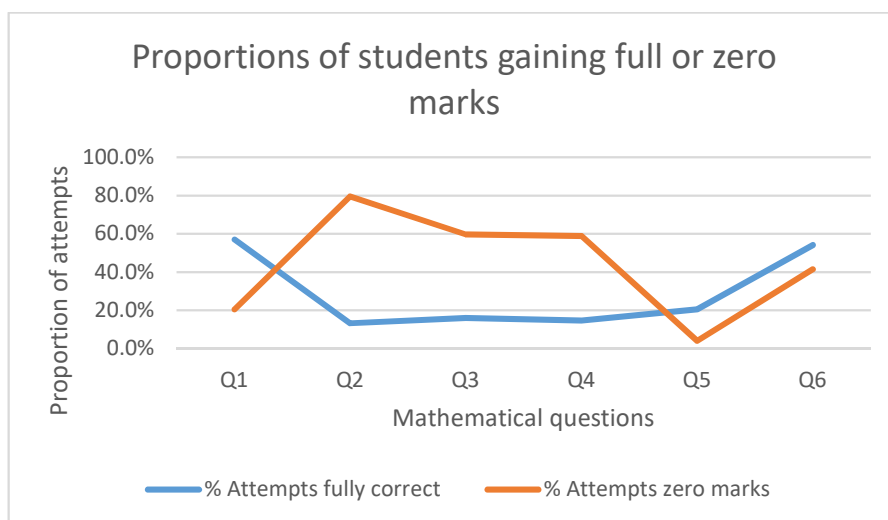
Source: Author's own.

From Table 12 it can be seen that not all the students attempted all the questions. 11 students (8.7%) did not attempt any questions at all; only 42 students (33.1%) attempted all six questions. The relative low levels of student engagement are not surprising, since the questionnaire was administered online and, for most students, there was no connection with the researcher.

For half of the questions (2, 3, and 4), the mark most commonly awarded to student answers was zero, whilst for questions 1 and 6 the most common award was full marks. The distribution of marks was polarised between the extremes of the mark range in questions 2 and 6, implying that these questions did not discriminate particularly well between students of differing abilities, whereas questions 1, 3, and 5 gained more of a spread of marks, suggesting that they were more discriminating; the reasons for these features are commented upon

further through the examination of students' responses to the individual questions. The relationship between the proportions of students who gained either full or zero marks against one another is shown in the line graph in Figure 26.

Figure 26 - Pilot Study: Proportion of Students gaining Full or Zero Marks



Source: Author's own.

Figure 26 reveals the extremes of the mark range clearly: for questions 1 to 4 there appears to be an inverse relationship between the proportions of students who gained zero marks and full marks. This might be predicted: it makes sense that, as higher proportions of students gain full marks, lower proportions gain no marks at all: that is, the distributions tend to be skewed either to the high end of the mark range (as in question 1), or to the low end (as in questions 2, 3, and 4). But for questions 5 and 6 this relationship is different: there is a direct relationship. In question 5, low proportions of students gained either zero or full marks (total: 24.7%), showing that the majority of the marks were distributed throughout the mark range, indicating that the question may have been a more discriminating one. For question 6, high proportions of students gained either zero or full marks (95.7% of students were included in one category or the other), showing how polarising this question was; analysis as to why is proffered in Section 5.3.

Table 13 - Pilot Study: Correlations between Students' Difficulty Estimates and Marks

Correlations between estimates and marks	Q1	Q2	Q3	Q4	Q5	Q6
Group P						
Correlation coefficient (<i>r</i>)	.10	-.06	.04	-.27	-.16	-.12
Sample size (<i>n</i>)	44	44	44	44	44	44
Significance (<i>p</i>)	.506	.954	.435	.785	.984	.902
Group Q						
Correlation coefficient (<i>r</i>)	-.45	-.19	-.09	-.46	-.44	-.40
Sample size (<i>n</i>)	83	83	83	83	83	83
Significance (<i>p</i>)	<.001	.081	.431	<.001	<.001	<.001

Source: Author's own.

Correlations were calculated between students' estimates and the marks they gained for answering the questions to establish the strength of the link. Pearson's product-moment correlation coefficient was used. Table 13 (above) shows these correlations.

No correlations were discovered that were particularly strong. It might be expected that there would be negative correlations (an inverse relationship) between students' estimates and the marks they gained – lower estimates would mean that students thought the questions were less difficult, and so they should have gained higher marks. In Group P, where students made their estimates of difficulty before attempting the questions, there were no correlations between students' estimates of difficulty and their marks. In Group Q, where students attempted the questions before offering estimates of difficulty, their marks and difficulty estimates were moderately⁶⁸ negatively correlated for Questions 1, 4, 5 and 6. The strongest link was found in Group Q, for Question 4, $r(81) = -.46, p < .001$.

⁶⁸ According to Field (2009, p. 173), values of more than $\pm .3$ can be said to represent a 'medium effect.'

5.2 Data analysis: inferential statistics

The use of inferential statistics enables analysis and exploration of tendencies within the estimates of the whole study sample, as well as differences between the two (P and Q) groups of students, and between male and female students. Overall, the total distribution of estimates was found to be left-leaning, indicating that there was a tendency for students to estimate questions as “easy” or “very easy” (39.5% of all responses), rather than “difficult” or “very difficult” (33.2% of responses). Given that, overall, the students tended not to score highly on the mathematics questions they answered, it can be suggested that they tended to under-estimate the difficulty of the questions. Within this overall distribution, however, there were some interesting differences between the P and Q groups, and between male and female students within those groups. Table 14 (below) shows the distribution of the estimates between students in the two groups. Whilst there are only small differences between the groups in the proportion of students who thought the questions were very easy or easy, it is noticeable that students in group Q had a marked tendency to judge questions more difficult or very difficult than students in group P, and much less tendency to be undecided.

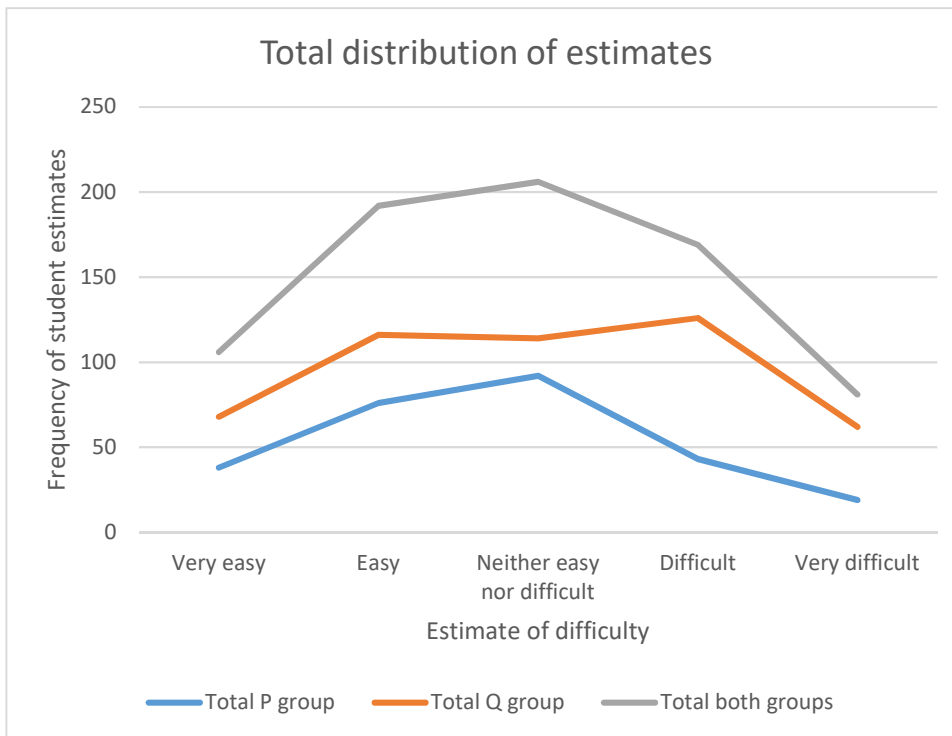
Table 14 - Pilot Study: Distribution of Estimates of Difficulty between Groups P and Q

Estimates of Difficulty	Very easy	Easy	Neither easy nor difficult		Very difficult	Total
			Difficult			
P group: Frequency	38	76	92	43	19	268
%	14.2%	28.4%	34.3%	16.0%	7.1%	
Q group: Frequency	68	116	114	126	62	486
%	14.0%	23.9%	23.5%	25.9%	12.8%	
Both groups: Frequency	106	192	206	169	81	754
%	14.1%	25.5%	27.3%	22.4%	10.7%	

Source: Author’s own.

Figure 27 (below) shows the same information graphically, using the number of estimates rather than the proportion, so as to separate visually the lines belonging to the different groups. The grey-green total line shows a slight left-leaning skew, with a higher hump on the left for the “easy” estimate than for the “difficult” estimate on the right. More interesting are the differences between the distributions for the P and Q groups.

Figure 27 - Line Graph Showing Distribution of Estimates between Groups P and Q



Source: Author’s own.

Two points present themselves: differences in the central tendency, and differences between the easy and difficult estimates. Focusing first on the number of students who opted for the central “neither easy nor difficult” estimate, there was a quite pronounced central tendency in group P (blue line), who made their predictions first, but this central option tended to be less popular with group Q (orange line), who answered the questions first. Looking at the differences between the numbers who estimated questions easy or difficult in each group on the difficult estimates, it can be seen that students in group P estimated questions easy rather

than difficult, and that students in group Q (who had already attempted to answer the questions) tended to estimate questions as relatively more difficult.

Bringing together this basic statistical analysis of marks, estimates and correlations, the perhaps unsurprising conclusion is that students who had already attempted the mathematical questions were able to estimate the difficulty levels in ways that correlated better with the marks they gained (as reported above, in Table 12). Students who had not already answered the questions tended to be more undecided on their difficulty estimates. This observation has consequences for the design of the main study that followed.

5.3 Discussion of pilot study survey questions

In accordance with the methodological considerations noted in Chapter 4, I made an assessment of the demands of the GCSE Mathematics questions used in the pilot and main studies, according to the CRAS Scales (Pollitt *et al.*, 2007). Table 14 (below) shows my assessment of the levels of demand for complexity, resources, abstractness, task strategy and response strategy. In order to gain a second opinion on this assessment from fellow professionals, I introduced the head of mathematics and another senior mathematics teacher at my school to the CRAS scales, and asked them to evaluate the question demands against the scale. After they had done this independently, we met and discussed our findings. There was a high level of agreement on most items, and as a result of the discussion I made some small adjustments to my assessments. As well as an assessed level (1-5), a brief explanation is also given for each type of demand in each question. The pattern of students' responses and explanations is subsequently analysed against these demands.

The GCSE mathematics questions that were selected for the pilot study were used again for the main study, with one additional question added. These questions therefore played a key role in shaping this project. Each question in the study is now discussed briefly to identify some of the issues of demand and difficulty that students encountered and that influenced their patterns of answers and estimates of difficulty. This question-by-question analysis provides the

start of our understanding of how students experience and comprehend demand and difficulty in examination question in mathematics.

In the discussion that follows, students are identified using the letter of the sub-group within the pilot study, P or Q, their entry number as allocated by the Microsoft Forms data collection software, and their sex. For example, P16F is student number 16 from the P sub-group, identifying as female; Q25M is student number 25 from the Q sub-group, identifying as male; and student Q114N is one who identified as non-binary or preferred not to identify their sex. Because the numbering system includes all responses collected, before the removal of those for which consent was not given, students in sub-group P are numbered 1 to 51, although there were 44 analysable responses, and in sub-group Q students are numbered 1 to 141, where there were 83 analysable responses. The total number of analysable responses was 127.

Table 15 - CRAS Scales of Demands for Pilot Study GCSE Mathematics Questions

Question	Complexity	Resources	Abstract-ness	Task Strategy	Response Strategy
1. Algebraic expressions	2: mostly single ideas and simple steps	1: all (and only) the data needed are given	2: mostly deals with concrete objects	4: students need to devise and apply their own strategy	3: some organisation of response needed
2. Probability	3: more complex and inter-dependent ideas	2: most of the information needed is given	4: mostly abstract	4: students need to devise their own strategy and monitor its application	2: organisation of response is straight-forward
3. Conditional probability	4: synthesis required; need to link cognitive operations	1: all (and only) the information needed is given	4: concrete concepts but abstract application	5: students must devise their own strategy without obvious checks	2: organisation of response is simple
4. Growth and decay	3: inter-linked steps needed	3: students need to select some of the information	4: mostly abstract	3: some strategy is given	2: organisation of response is straight-forward
5. Algebraic expressions	3: mostly single steps but more complex for last part	3: information is provided but needs selection	4: mostly abstract	3: some strategy is given, but some must be provided	3: some organisation required of response
6. Mensuration	2: single ideas and simple steps	1: all (and only) the data is given	2: mostly concrete (circle)	3: students expected to remember and apply familiar formula	1: organisation of answer is very simple

Source: Author's assessment (following peer consultation) of pilot study question demands, applying CRAS Scales from Pollitt *et al.*, 2007.

Question 1

Reuben hires a car.

It costs £150m plus 85p for each mile he travels.

When Reuben hires the car, its mileage is 27,612 miles.

When Reuben returns the car, its mileage is 28,361 miles.

How much did Reuben pay to hire the car? (4 marks)

Question 1 concerned algebraic expressions. It required a multi-stage approach, with arithmetic that is straightforward to apply on a calculator. Complicating factors that added to the cognitive load and demand were that students needed to create their methods, and they also needed to convert their units from pence to pounds. The question required students to devise, apply, and monitor their own task strategies. First, they needed to work out how many miles Reuben's hired car had travelled, by subtracting the car's starting mileage (27,612) from its mileage on return (28,361). They then needed to multiply the number of miles travelled by the cost per mile (85p). Finally, they needed to add the fixed hire cost of £150 to the result of their multiplication. There were opportunities to make mistakes at each stage, but the mark scheme allowed method marks to be awarded for each one of the three necessary steps.

With a mean value of 2.7, most students considered this to be one of the easier questions.

CRAS scales (Table 15) rated this fairly easy, particularly with regard to complexity, resources, and abstractness. Students' estimates, which suggested that this was one of the easier questions, mirrored the comments from the examiners' report:

'There were some excellent responses to this question, showing all the required steps, dealing with units of pounds and pence correctly and reaching a correct solution. The majority of candidates made an attempt at this question, with many scoring at least 1 mark for attempting to calculate the distance covered. Multiplication by cost per mile was usually applied correctly using £0.85 or 85p' (OCR, 2018c, p. 14).

Looking at student answers to this question, 84.3% of respondents gave an answer, 61 of which (57.0%) were fully correct. This ranked the question highest for the marks that students secured. On this empirical measure of difficulty, therefore, this question was the least difficult on the questionnaire. The question appeared in the first half of the foundation tier paper,

which indicates that examiners also considered this question to be relatively straightforward.

Examiners noted some complicating factors and features of the question that caused some students to stumble:

‘If marks were subsequently lost it was often due to not equating the units of pence and pounds when adding to the standing charge of £150. Many simply did not add at all and stopped at a value obtained from their multiplication (636.65 if calculated correctly). There were a number of candidates who did not check that their answer was reasonable in the context of the question and offered unreasonable amounts for the cost of hire’ (OCR, 2018c, p. 14).

14 students (11.0%) failed the ‘reasonableness’ tests mentioned by the examiners’ report, entering a hire price of many thousands of pounds: £63,815 was the most common error.

There was no significant correlation between estimates of the difficulty of the question and students’ actual performance in the question for students in Group P, although for students in Group Q their marks and estimates of difficulty were negatively correlated, $r(81) = -.45$, $p < .001$. Of the 20 students overall who considered the question to be very easy, 2 gave no answer, 3 scored 1 or 0 marks, and the other 15 scored full marks for their answers.

Conversely, of the 3 students who estimated that this question would be very difficult 2 scored full marks and the other did not attempt an answer.

Two opposite tendencies can be seen in students’ responses to this question, which follow through into subsequent questions, concerning whether a question was considered to be easy. This theme is developed in more detail in Section 5.4. Some of the students appeared to rate the question easy despite the fact that, in practice, they did not tackle it. This would suggest that, when they initially looked at the question, they did not understand the demands that the question would make of them. They may have had little understanding of their own learning and problem-solving processes.

One student (P22F) estimated that the question would be ‘easy,’ adding that it was ‘*quite easy and straight forward*’. She gave the most succinct answer, showing her workings, and gained full marks:

$28361 - 27612 = 749$
 $0.85 \times 749 = 636.65$
 $£636.65 + 150 = £786.65$

The reasons that students gave for their estimates can be analysed into three main categories; 7 responses did not fall into any of these categories. The first and largest category (67 students, or 60.4%) gave answers that showed that they felt at ease with the question. The next largest group was comprised of 32 students (28.8%) who felt confused. The final 12 (10.8%) analysed the question more deeply. Of the 7 outliers, one gave a relevant numeric calculation as a response, and the others gave variants of “neither easy nor difficult” but gave no further expansion of their view.

24 students from group P observed that the question required only ‘*simple maths*’ (P21M), or that they were ‘*able to use my knowledge*’ (P42F) to tackle the question. In the other group, student Q06F wrote for this question, ‘*multiple steps but from skills from earlier in school*’. One student (P44M) noted that the arithmetic would be ‘*especially easy on a calculator paper*,’ which this was. Of these 24 students, 6 rated the question very easy, 12 rated it easy and the remaining 6 neither easy nor difficult. One student, who estimated that the question was very easy, explained the method succinctly:

‘Because all you have to do is take away the first mileage away from the second mileage and multiply it by 85 then add it to 150’ (P38F).

She did not mention the differing units, however, (85p per mile, but £150 for the fixed hire charge), and in fact she did not subsequently enter an answer. Of the students who rated the question very easy or easy, most gained full marks; 2 did not enter an answer; only 1 did not score any marks.

32 students felt “confused” or said that they ‘*don’t know where to start*’ (P25F). Looking more closely at these 32 students’ responses it can be seen that their estimates mirrored their confusion: although 9 rated the question neither easy nor difficult, the remainder rated it difficult or very difficult. 2 of them also said that they would need to use ‘*trial and error*’ (P9M;

P17M), although it is not clear that this question did, in fact, lend itself well to such an approach.

Eight students explained that there were several different steps to carry out to reach the answer. They showed differing levels of insight into how and whether the presence of more steps made the question harder. One wrote that, *'You have to do working out in four different parts but I think I could do it with time'* (P32F). This student was correct in her estimate of her own ability to solve what she estimated to be a difficult question, and in her answer she gained full marks.

One student explained that, *'I think it is difficult because they are a number of different steps in order to get the answer'* (P36F); another wrote that *'it was difficult because there are to [sic.] many steps and you have to figure out so many things for one question'* (Q47M). These students equated question difficulty with the multiple steps that needed to be taken but made no comment on the demands posed by any of the individual steps. These students articulated at least some understanding of the complexity of the question, where it was the formulation of the method rather than the standard of the arithmetic that increased the level of demand in the question. One student, who rated the question very difficult and who explained, *'i didn't know where to start'* (P25F), nonetheless gained full marks when she did tackle the question. This suggests that although the student's self-efficacy and self-belief may have been quite low, once her cognitive system was engaged, she was equal to the task.

12 students were insightful about the steps needed to tackle the question, and about the possible trip hazards and opportunities for error that the question posed. Most of these students rated the question easy or very easy, but 4 rated the question difficult, 1 rated it neither easy nor difficult, and 1 very difficult. One student understood the question requirements clearly, explaining that:

'The question is easy to understand, and it seems pretty simple what you need to work out. However, some people may forget to add on the £150 at the end, after calculating the mileage cost' (P47M).

A student who rated the question very difficult but who subsequently gained full marks, articulated a degree of ambivalence around how demanding the question might be. He explained some possible sources of error as well as the necessary method:

'The question isn't too difficult because of how straightforward the solution is to solve the problem. It also isn't too easy since there is significant room for error. For example, you could misread the question and multiply the 85 p by 28,361 miles (instead of subtracting the mileage when from when he returns the car from when he hires it), you may also mistakenly think you have to multiply the cost of £150 by the amount of miles he travels. There is enough information in the question that you need to pay attention to, but not so much that you would feel overwhelmed' (P48M).

This was an interesting response, and indicated some awareness of his own cognition and mathematical thinking. The student began by making an unacknowledged link between his own ability and the question's demands (*'it isn't too difficult,'* he noted because, essentially, he believed that he could solve the problem). At this point, therefore, the student appeared to be 'consciously incompetent' to borrow the terminology from the Noel Burch model discussed in Chapter 3: he didn't really know whether the problem was difficult or not, since he had not thought his way through it; he assumed it would be straightforward. However, he then started to unpack the sources of possible error, and as he did so he developed his understanding, moving into a state of being 'consciously competent,' and describing several traps that did in fact befall his fellow students in this survey. Student P48M also exhibited high levels of self-efficacy: he believed he would be able to solve the problem, even before he started to tackle it in detail: his phrase *'it isn't too difficult'* shows self-confidence. Although many students considered Question 1 to be straightforward, then, their responses and misconception have displayed some of the themes and trends that will be explored further in other questions.

Question 2

Finn has two bags of counters.

He takes a counter at random from each bag.

The probability that he takes a red counter from the first bag is 0.3

The probability that he takes a red counter from the second bag is 0.4

What is the probability that he takes at least one red counter? (4 marks)

Question 2 caused problems for many students; the examiners' pithy comment, that 'there were very few fully correct answers' (OCR, 2018c, p. 20) showed that this was also the case when the question was set in the 2018 examination. "Easy" was the most frequent estimate of difficulty; 98 students (77.1%) gave answers but it attracted the smallest proportion of fully correct answers (13.3%), and 79.6% of answers gained 0 marks. Application of the CRAS scales (Table 15) shows that most of the demand in the question lay in the abstractness of ideas and the need for students to devise their own task strategies.

This question was not a good discriminator: 92.9% of students who entered an answer gained either zero marks or full marks. The mark scheme allowed for 3 method marks to be given for a very specific incomplete answer, but none of the students in this study offered that answer.

What was remarkable about this question, from the perspective of this study, was that so many students, including many who had already attempted to answer the question, were so wrong in their estimates of its difficulty. Looking at group Q, who made estimates of difficulty after attempting the question, 35 students (50.0%) subsequently rated it as very easy (12) or easy (23); 24 of these students gained no marks. A further 18 (25.7%) also failed to spot the question's difficulty, estimating it to be neither easy nor difficult; of these, 15 gained no marks. In total, then, three quarters of students who had already attempted the question underestimated its difficulty, and three quarters of those students gained no marks.

Across the whole of the pilot study sample, 23 students (18.1%) expressed uncertainty or explained one or more aspects of difficulty, without gaining any marks. One of students noted,

'This question requires you to draw a probability tree, which can be quite tricky sometimes as you need to think about what to put on each of the branches, and usually there would be 2 different colours - not 2 different bags - which makes this question that little bit harder' (P47M).

This starts to open up some misconceptions. The question referred to drawing a red counter out of two bags. The possible colour of the other counters was irrelevant. Another student appeared to have been similarly confused by the two bags, stating for her answer, '0.7 out of 2.0' (P25F).

Some students became confused about when to use the addition and multiplication laws of probability. One commented that *'because probabilities add up to 1 so you would add the numbers then take away from 1'* (P20M). His idea about all the probabilities adding up to 1 is correct, but he should have multiplied the probabilities first, before subtracting from 1.

Another student started well enough, and stated that *'the first bag of red is 0.3 but then putting it back the probability for red is 0.3 and not red is 0.7'*, before confusing the issue: *'but if he takes a not red it's 0.7 but then puts it back is red is 0.3 and not red is 0.7'* (P34M). This student was perhaps thinking of the act of taking a counter, then putting it back before taking out another counter when, in reality, the probability of the event (taking a red counter) stayed the same, because the events were independent of one another. He had misunderstood what this question was asking. This is an interesting point – the student's response showed that the source of difficulty for him was not the mathematics of probability, but his inability to decipher the question to see what the problem actually was.

These students presented answers that suggest that the question may have been inappropriately focused. Ahmed and Pollitt noted that,

'A context in which the questions are focused will help to activate relevant concepts in the students' minds and so be less likely to cause them to form a mental representation of the question different from the one the examiner intended' (2007, pp. 205-6).

Here, some students appeared to be led to expect a different question from the one that was posed: *'it did not mention another colour and its a question I am not used to'* (Q47M).

There are implications here for examiners and possibly also teachers. This question divided students, but it did not discriminate well: students gained either full marks or no marks. There was no meaningful correlation between estimates of difficulty and marks gained. Of those who gained no marks or who did not answer, there was a wide spread of opinions about the difficulty of the question. Although this was a foundation tier question, the comments in this study and the examiners' remarks from 2018 show that most students did not understand clearly how to solve the problem. This question appears to have been more difficult in practice than the examiners had expected and predicted, possibly owing to the insufficiently focused formulation of the question. Had the examination questions been pre-tested and their performance analysed, the poor fit of this question might have made it less likely to have been included in an examination, at least in the form in which it was presented. In the classroom, perhaps some students need to experience questions in a variety of different contexts and formats, so that they are not misled into wrongly predicting or presuming what a question might be asking.

Question 3

**60% of the people in a town are males.
20% of the males are left-handed.
21.6% of all the people are left-handed.**

Work out the percentage of the people who are not male who are left-handed. (5 marks)

Question 3 was attempted by 87 students (68.5%), the second-lowest percentage recorded for any of the six questions. Only 14 (16.1%) were fully correct. 52 (59.8%) scored 0 marks. The mean mark was 1.3 and marks were more widely spread for this question than for any others (standard deviation = 1.9). According to the application of the CRAS scales (Table 15), this was

one of the most demanding questions: although all the information needed was supplied, it was a complex abstract problem that required students to devise a complete task strategy.

In group P, 4 students (9.1%) just stated that it was easy; another ten (22.7%) explained why they thought it was easy. In group Q, 8 students judged the question to be easy. Most of these responses were of the nature either that the question was “easy to understand”, or that it offered “simple working out”. The cognitive load, in these students’ estimations, was not heavy. However, of these 22 students, 12 gained no marks; 8 did not attempt an answer. These students could be described as unconsciously incompetent: they thought the question was simple, but they did not recognise that they did not know how to answer it.

Across the whole study, 4 students (4.5%) who described the question as very easy or easy gained full marks. These were ‘consciously competent’ students. One stated that it was, *‘quite simple, in my opinion. I’ve done this type of question numerous times’* (P27F), making the point that repeated practice made the question simple for her. This idea of repeated practice leading to deeper learning is embedded in mathematics education, and is supported by findings in neuroscience. Hohnen and Murphy (2016, p. 79) refer to repetition or practice that results in what they term the ‘myelination of that circuit,’ where neural networks in the brain work more efficiently after repeated practice. One of the aims of the mathematics programmes of study in the National Curriculum in England (DfE, 2014b, p. 3, emphasis added) is that, ‘All pupils should become fluent in the fundamentals of mathematics, including through varied and frequent practice, so that pupils develop conceptual understanding and are able to recall and apply their knowledge rapidly and accurately to problems.’

In contrast to the P group, students in Q group (who had already attempted the question before giving their estimations) tended to estimate the question as more difficult. Of the 25 students who did not give an answer to the question, 10 rated it difficult, and 13 very difficult.

Students' reasons for their choice of estimate showed they did not understand the question or how to tackle it. 39 students in group Q (47%) said they "did not understand" (or some variant) or were "confused" by the question. Contextual and presentation factors affected some: *'the context of this question confused me'*, wrote one (Q44M); another stated that the *'layout/ explanation is a bit complicated'* (Q62F). One claimed that it was *'just a difficult question. could be worded clearer'* (Q42M), whilst another suggested that *'I think this would've involved a Venn diagram but wasn't sure how to answer it'* (Q23F), and another also wondered about a Venn diagram but gave no details about how it might be used.

Examiners expected not a Venn diagram but a different organisational method: their report states that *'the best solutions either used a tree diagram or they started with a population of 100 and produced a two-way table'* (OCR, 2018d, p. 18). This was a question where the electronic format of the questionnaire did not allow the researcher to see students' trial approaches to any diagrammatic solution: there was no mechanism by which the students could draw tree diagrams, two-way tables, or Venn diagrams in their electronic answers. In the absence of being able to see students' written attempts or discuss their thoughts with them, it is only possible to look at any workings or incorrect responses entered onto the questionnaire. Of the 38 students in group Q who entered an incorrect answer, 24 (63%) supplied no further explanation or workings, making it impossible to probe for misconceptions.

The most common incorrect answer was caused by misunderstanding what the question was actually asking for. The instruction was to,

'Work out the percentage of the people who are not male who are left-handed.'

This appears to have confused many students, and to have been widely misinterpreted by others. Pollitt *et al.*, have observed that, in the context of the wording of examination questions, that, *'the most difficult word in the English language is probably "not"'* (2007, p. 194). Had the examiners drawn students' attention to the occurrence of word 'not' in the question, then the proportion able to answer correctly might have been increased.

Several students got as far as calculating the proportion of the total population that was both not male and left-handed, which came to 9.6%. However, this is not the final answer to the question: examiners sought ‘the percentage of the people who are not male who are left-handed.’ Students were therefore expected to relate the 9.6% of the total population back to the proportion of ‘the people who are not male’ (which is 40%). Those who spotted this relationship needed to divide 9.6% by 40% to get the fully correct answer of 24%. Even some students who were able to answer fully correctly admitted some difficulties with this question. Student Q131M wrote that he had been ‘*unsure how to link the full population to the male left handers;*’ another was ‘*not sure what I’m meant to do*’ (Q57M); a third reported that the ‘*layout/ explanation is a bit complicated*’ (Q62F), and a fourth (Q108F) was just ‘*not sure.*’ These students all rated the question as difficult. Only one student, who rated it neither easy nor difficult, noted that she ‘*thought the steps were pretty basic and easy to understand*’ (Q21F) and answered correctly. Again, this question brought up the issue of focus; discussed further in Chapter 7.

Question 4

The value of a car, £V, is given by

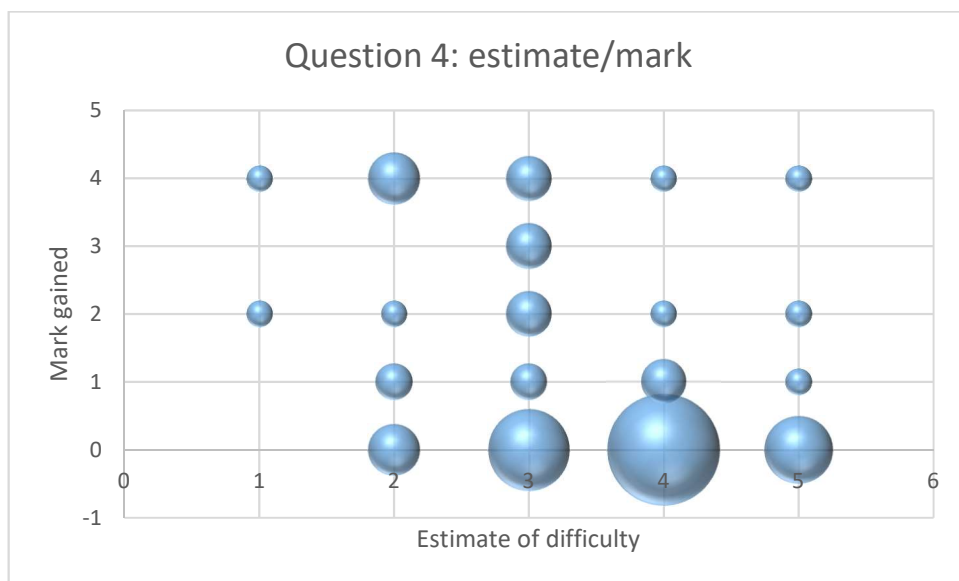
$$V = 16,500 \times 0.82^n$$

where n is the number of years after it is bought from new.

- Write down the value of the car when new (1 mark)
- Write down the annual percentage decrease in the value of the car (1 mark)
- Show that the value of the car after 4 years is less than half its value when new (2 marks)

Question 4 was about growth and decay. Many students appear to have been daunted by aspects of this question: only 68 out of 127 (53.5%) attempted an answer, the lowest for any of the six questions. The application of the CRAS Scales (Table 15) indicates that the demanding factors were that this was a mostly abstract problem that needed some selection of information, the application of the method given, and the construction of a task strategy for the final part. In terms of estimates, difficult was the most frequent (36.2%), and students rated this question the most difficult (mean = 3.7) with the smallest spread (standard deviation = 1.0). However, the question was more discriminating than most, with students accessing the whole range of marks, from 0 to 5.

Figure 28 - Bubble Chart Showing Frequency of Estimates and Marks for Question 4



Note: The diameter of the bubbles indicates the size of the frequency for each pair of estimate/mark points: a larger bubble shows a higher frequency.

Source: Author's own.

The bubble chart graphic in Figure 28 highlights that, of the students who gained 0 marks, more of them rated the question either easy or neither easy nor difficult than the number who rated it very difficult. 19 students rated the question difficult and gained 0 marks. Students who scored high marks, and those who scored low marks, came from the whole range of difficulty estimates. There was no correlation for students in Group P between estimates and marks, but for students in Group Q there was the highest (negative) correlation between estimates and marks for this question: $r(81) = -.46, p < .001$.

Similarly to Question 3, there were some presentational issues that created difficulty for some students. Even some from Group Q who were more successful in their answering strategies admitted to experiencing some difficulties: one stated that it *'wasn't as clear as some of the other questions'* (Q32M). Two other successful students wrote to similar effect: one, who graded the question neither easy nor difficult stated that, *'the question was fairly simple however i had to think about this one a bit more'* (Q02F); another wrote, *'i find percentage increase/decrease easier but the first question i had to think more'* (Q40M). This latter student graded the question easy and gained full marks; it would have been interesting to hear from all these students what aspects of the question were not clear or made them need to think more, but none of them gave any further details. There are hints of metacognition awareness here, as students made reference to needing to think through the question, but they are not explicit, so definite conclusions cannot be drawn as to these students' awareness of their own question answering processes.

Question 5

Solve

a) $4x = 56$ (1 mark)

b) $8x - 6 = 46$ (2 marks)

Solve by factorising

c) $x^2 + 11x + 30 = 0$ (3 marks)

Question 5, like Question 1, was about algebraic expressions. Like Question 4, it proved more discriminating, with marks gained across the whole range; challenges were presented in the different sections of this multi-part question. CRAS Scales (Table 15) indicated that the question posed a number of demands, almost all at a moderate level, but increasing for the final part of the question. It was rated one of the more straightforward questions by students, with 62 students (48.8%) rating it very easy or easy. The opinion of the students was borne out by their marks; few students (only 4, or 4.1%) scored zero marks; 20 (20.6%) gained full marks; students gained marks for different parts of the question and consequently scored across the mark range: 67 (69.1%) gained half marks or more. This question was out of 6 marks; the mean was the highest for any question, at 3.0, and the spread was also relatively wide (standard deviation = 1.8).

This question was in the first part of the foundation tier paper (number 7 of 20 questions), which would imply that examiners considered it, overall, to be relatively straightforward. However, this proved to be a question with widely varying levels of difficulty in its different parts. Examiners' comments explained that,

[The first part of the question] 'did not cause many problems as only one operation was required to obtain the answer...'

[The second part of the question] 'was well answered by many candidates. Those who were not able to arrive at a correct response of 6.5 rarely used algebra and consequently lost method marks. The most frequent error was to give an answer of 5 from $(46 - 6) \div 8$ but, unless an algebraic method was shown, this was unlikely to score' (OCR, 2018c, pp. 10-11).

Answers given by students in this study bear out the examiners' remarks: almost all students gained marks in the first parts of the question. The examiners also noted that the last part of the question was different and, they infer, more demanding:

'Only a minority of candidates understood the requirements for answering this question. Of those who did, several gave the roots as 6 and 5 failing to realise the significance of $x + 5 = 0$ and $x + 6 = 0$ ' (OCR, 2018c, p. 11).

Two thirds of students (65 students) gained between 1 and 3 marks for this question, correctly answering the first two parts of the question. Students found the third part considerably harder. Here the complexity was caused by students needing to be able to handle formulae and equations in algebra confidently, and to have practised factorisation frequently enough to be fluent in applying the method without prompting or assistance. The results obtained suggest that students in this survey tended to over-estimate their own ability in this area: most thought they could handle algebra and factorisation, but few scored well.

97 students attempted this question, making it one of the questions that most students answered. This possibly reflected the estimates that they had given; only 24 students (18.9%) judged the question to be difficult or very difficult. The picture that emerges from this study in this regard is that if students judge a question to be straightforward, then they are more willing to have a go at answering it. Conversely, students are put off attempting questions which they perceive to be difficult. This is an important effect of students' confidence and self-efficacy: if they believe a question is possible, they are willing to have a go, but they are less willing to risk failure.

In group P, where students had not attempted the question when they made their estimates of difficulty, a smaller number of students than for previous questions (5, or 11.4%) expressed some doubt about their ability to meet the demands of the question. One noted that it was *'not easy to work out'* (P23F); another wrote that it would *'take a little bit of time'* (P17M); a third reckoned that *'I can do the first half but would need help for the second half'* (P14F).

These students all gained between 0 and 3 marks. In contrast, students from group Q, who had already had an opportunity to answer the question before reporting on its difficulty, tended to be more negative in their estimates, particularly of the last part of the question: *'A. and B. were easier but C. was a bit more challenging because I couldn't remember how to do it at first'*, wrote student Q27F; these sentiments were echoed by student Q39M (from a different school), who wrote that *'I found the first two quiet [sic.] simple as I have been learning these in*

lesson, but I struggled on the last one due to me forgetting how to do the question'. From this latter school, another student (Q41F) also referred to the work they had done in lessons: 'We practice equations and quadratics a lot and I like them'.

Repeated practice was a factor mentioned by other students: some referred to having done 'much practice on it' (P10M) or stated that it was 'simple, done many times before in class' (P27F). In all, 15 students (11.8%) referred to "practice" or having done questions like this before. Confident language choices again demonstrated high levels of self-belief – and by implication also indicated familiarity and practice with this type of question; for example: 'I know what I am doing and all you have to do is simplify it or factorise it' (P38F), and 'you just have to do the same thing to both sides and know your times tables to work it out' (P35M). 21 out of 24 of these group P students, however, overestimated their ability, or underestimated the difficulty of the last part of the question; most gained the first 3 marks out of the total of 6 available. Question 5 was a good example of a discriminating multi-part question that, because challenge (= demand) increased through the different parts, was able to give all students an opportunity to succeed, whilst also stretching and testing those who were more expert.

Question 6

A circle has radius 6cm

Calculate its circumference

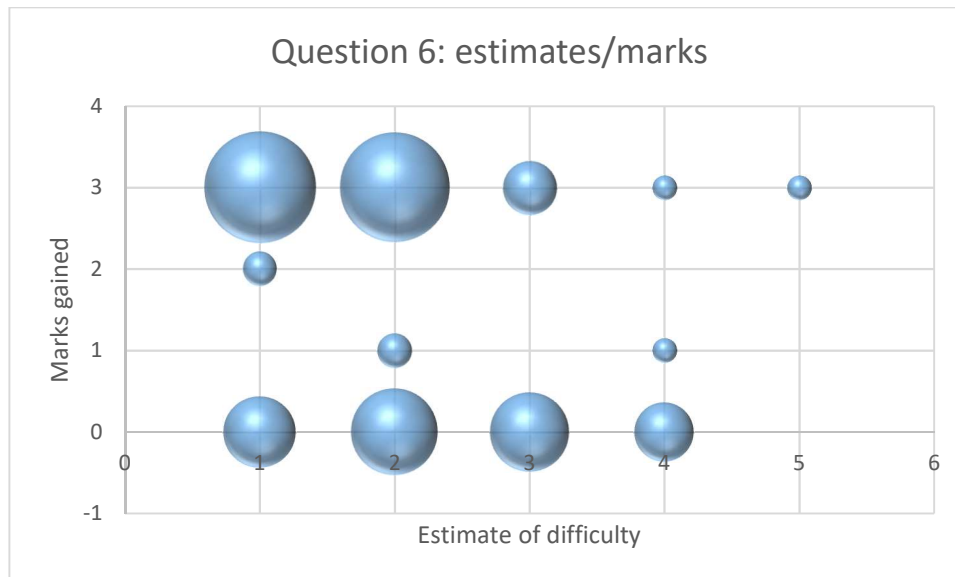
Give your answer in centimetres, correct to 1 decimal place (3 marks)

Question 6 appeared to be straightforward to students, and they rated it the least difficult (mean = 2.4); 76 students (59.8%) rated it very easy or easy. The application of CRAS Scales (Table 15) showed that this was the most straightforward question, the only significant demand being the requirement to remember and apply the correct formula for calculating circumference, a necessary skill for the mensuration part of the GCSE mathematics course. The question did not test problem-solving skills; it tested accuracy of recall and the application of a familiar formula, which is an important skill for a mathematician. The question simply distinguished between students who knew the formula and could apply it correctly (39

students, or 54.3%), and those who did not remember it or/and could not apply it (35 students, or 41.5%). This explains an apparently anomalous situation, where the proportions gaining zero marks and full marks were both high.

Students who did not remember the formula to calculate the circumference of a circle, or who confused diameter and radius, therefore did not score well in this question. The presumption might be that students who thought this question was easy or very easy would be those who did recall the formula, and who should therefore have gained full marks, and those who ranked it difficult could not remember and gained 0 marks. The bubble chart in Figure 29 shows that the situation is not so simple, however. 69 students (73.4% of those who attempted the question) rated the question very easy or easy. Although the majority of these students (43 out of the 69, or 62.3%) gained full marks,⁶⁹ a sizeable minority (22 students, or 31.9%) gained 0 marks.

Figure 29 - Bubble Chart Showing Frequency of Estimates and Marks for Question 6



Source: Author's own.

⁶⁹ Two students who estimated the question as difficult or very difficult also gained full marks: of these, one gave no reason; the other correctly stated a version of the formula: ' $C=2\pi R$ ' (Q18M).

Looking at the 22 students who thought that the question was easy or very easy but gained 0 marks, it can be seen that their answers showed that they had not accurately remembered (or applied) the formula. This question required students to have prior knowledge of the formula for calculating circumference and, having remembered it, to apply it correctly. The correct formula was $2\pi r$ or πd , but some misbegotten variants were used, such as r^2 or $2r$. Most common was the use of the formula for the area of a circle (πr^2), where circumference was what was sought by the question. It is not possible to tell whether students confused the meanings of the terms circumference and area (there was no significant indication of this, however, and although one student, Q34M, wrote *'I have never understood radius and circumference'* he does not mention area), but it seems most likely – particularly given the other variants used – that students simply misremembered the formula. One claimed that she had *'remembered the formula to work out the circumference of a circle'* (student Q33F) – but she had not – and two others helpfully wrote out the error: *'pi times the radius squared'* (Q47M and Q57M). Examiners noted in their post-examination report (OCR, 2018c) that some students had confused the formulae, but it is not clear from their comments whether they had in any way anticipated that this might be so.

Students who thought the question was difficult tended to give answers showing that they *'don't understand circumference'* (Q112M), or had *'forgotten how to work out circumference'* (Q109M). One admitted that *'I was learning this a lot during lockdown⁷⁰ but it does get confusing remembering the methods for all the different ways to do with these type of questions'* (Q48F).

In some earlier questions (particularly questions 2, 3, and 4), students admitted to being confused. But when a student who expects a question to be easy either mistakes or

⁷⁰ The lockdown, to which this student was referring, was the lockdown between January and March 2021 when schools were closed owing to the coronavirus pandemic, when most students had to learn remotely via online learning, without the opportunity for students to check their understanding regularly or for teachers to identify misconceptions.

underestimates its demands, they are deluded or misled rather than confused. In terms of the Burch competence model, this puts them into the 'unconscious incompetent' quadrant. Their initial rapid judgement, which, as was seen in Chapter 3, may precede the engagement of their cognitive systems, has let them down. The consequences of this for a student in an examination could be substantial: if they have not sufficiently or accurately appraised the demand of a question and they have under-rated its difficulty, they are unlikely to address it adequately.

As students' responses to these six mathematical questions have been discussed, many of their preoccupations and thoughts have been uncovered, telling the story in their own words. Now it is time for these responses to be brought together systematically in the thematic analysis. The complicated narrative of the data will be condensed in a way that attempts to address the research question and also 'provides a concise, coherent, logical, non-repetitive and interesting account of the story the data tell' (Braun and Clarke, 2006, p. 93). Given the relatively low levels of engagement of most students in this survey, as shown by the short amounts of time they spent on it, the responses available for reflexive thematic analysis were not of the richness that might have been hoped for. This factor accounts for the rather semantic (surface meaning) level of analysis that follows for the pilot study.

5.4 Reflexive thematic analysis

When first coding students' responses, a large number of possible areas were identified. These putative themes included students' reactions to questions, such as "easy" or "confusing"; their recognition and description of "different steps", building towards a "method"; their "confidence"; and the ways in which "prior learning" and "repeated practice" contributed towards their understanding and ability to address a question. There was then an extensive period of reading and re-reading, relating the data to one another in different ways, mapping possible themes and recoding. A worry was that the themes were too large, too indistinct and

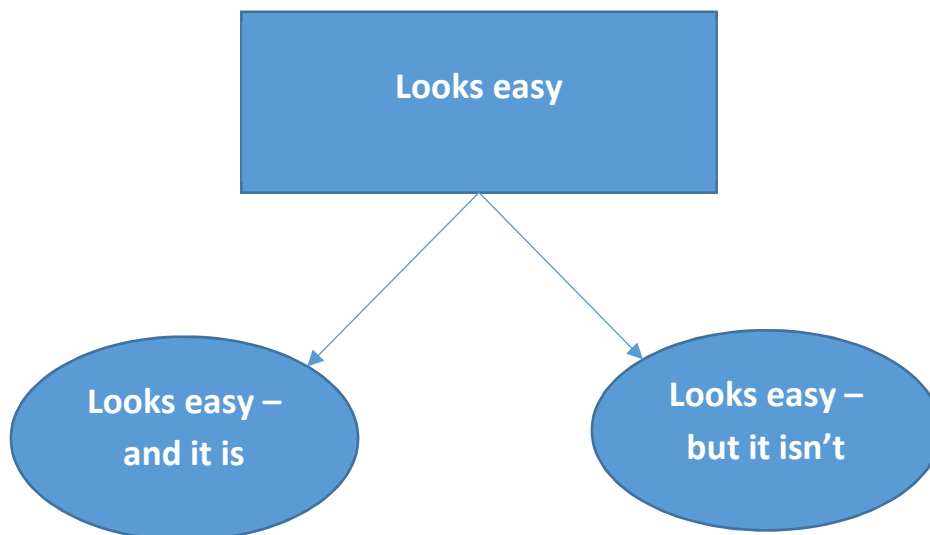
not coherent enough. This was perhaps partly because the levels of student engagement – through the online method of data collection mandated by the post-COVID restriction – were not as deep as had been hoped for. Through a thorough review of these themes, using a reflexive process and relating the developing themes back to the theoretical underpinning of the study, it became clear that some of the data were not relevant, not strong or coherent enough to stand as themes, while others were developed into sub-themes, nested within more coherent and defined themes. This process of theme development at times felt tortuous for this part of the study, a feeling familiar to many involved in reflexive thematic analysis (see Braun and Clarke, 2022, pp. 102-112).

The two themes have been named as “looks easy” and “recognising steps”. These themes have been named after the semantic, surface level of the data, but in both cases the meaning and implications of the themes lies at a deeper, latent level, in terms of what they reveal about students’ engagement with examination questions and where the sources of difficulty lie for them. Following the identification and development of these two main themes, a further period of engagement with the whole dataset was undertaken, to see if perspectives had changed or that different nuances were noticed in the data. A point that became increasingly important at this stage was the apparent contradiction within the theme “looks easy,” in that student responses “pulled” the theme in two apparently opposite directions. Many students appeared to feel that a question looked easy, but it became apparent on analysis, comparing their estimates of ‘expected difficulty’ with their marks (the outcome of ‘experienced difficulty’) that they had not understood the demands of the question. On the other hand, another set of students evidently had the understanding of how to tackle a question successfully to back up their estimation that the question “looks easy.” Each of these themes is now explored in turn.

5.4.1 “Looks Easy” theme

With regards to the looks easy theme, it is evident that students’ emotional responses were engaged, as well as their rational responses. This emotional-rational dichotomy is an important distinction and relates to the theories of Bandura (1986), Marzano and Kendall (2007), and Burch’s competence model; addressed in more detail in Chapter 7. When a student first encounters an examination question, they often seem to make a rapid judgement: either it looks easy or it does not. If the student thinks it “looks easy”, this may well mean, “I think I can do this.” This judgement is a prediction of success; however, it may not relate well to the student’s subsequent actual success. The “looks easy” theme therefore has two sub-themes: “looks easy – and it is”, and its opposite, “looks easy – but it isn’t”. The theme is illustrated in Figure 30.

Figure 30 - “Looks Easy” Theme



Source: Author’s own.

With regards to the first sub-theme – “Looks easy – and it is,” patterns are evident in what the students wrote, and a theme was developed. Students’ responses around questions 3 and 4, which most estimated (and found) to be more difficult, produced particularly pertinent comments. Confidence was a hallmark of many of the students’ responses: one explained that she thought it was easy *‘because I understand the question and will be able to give a good answer’* (P30F); another had analysed the question and wrote that it was, *‘just compound interest’* (Q130M), so clearly not something that worried him; student Q57M felt there was a *‘logical conclusion to the question’*. Several students who recognised the demands of the questions and were able to meet them gave comments that, although they fall short of describing methods, gave insights into their thinking: student Q128M noted that question 3 *‘was slightly harder but still easy as it was just about making the percentages real to a certain number of people’*, suggesting some visual modelling in his approach. Student Q46M suggested, *‘just find out how many males the 20 percent include the[n] add the rest of the percentage’*. The word ‘just’ is significant in both of these responses – it infers a confident lightness of touch. Student P37F asserted that she was *‘pretty confidence [sic.] with percentages’*. Students in the two groups, P and Q, gave answers that were similar. As noticed earlier in this chapter, however, fewer students who had already attempted the questions rated them as easy, showing that students are less able to give an accurate rating of the demands of questions they have not yet attempted.

In exploring the “Looks easy – but it isn’t” sub-theme, responses of the two groups of students will be examined separately, because they reflect different nuances: students in group P predicted that questions would be easy but subsequently found that they could not meet their demands; students in group Q had already attempted the questions – and had not given the correct answer – but still thought they were easy. Some students in group P who over-estimated their abilities in relation to question demands, showed that they were over-confident. Student P05F, for example, stated that she was *‘able to understand’* the requirements (for three questions which she got wrong). Student P28F asserted, *‘I’m good at*

working with probabilities,' and P29F wrote, *'I find these type of question easy'*; neither was successful. Other students gave sketchy and over-optimistic accounts of their methods which did not survive their encounters with the question: student P33F wrote, of question 1, that you *'just have to subtract miles and multiply with the price'*; student P45M confidently stated, for question 2, *'I find probability trees to be something I can handle so this shouldn't be difficult'*. Stereotypically, and from existent research (for example, Parker *et al.*, 2018), it might be expected that this slapdash tendency would be more prevalent among male students, but in the present study this approach was seen almost equally among female and male students.

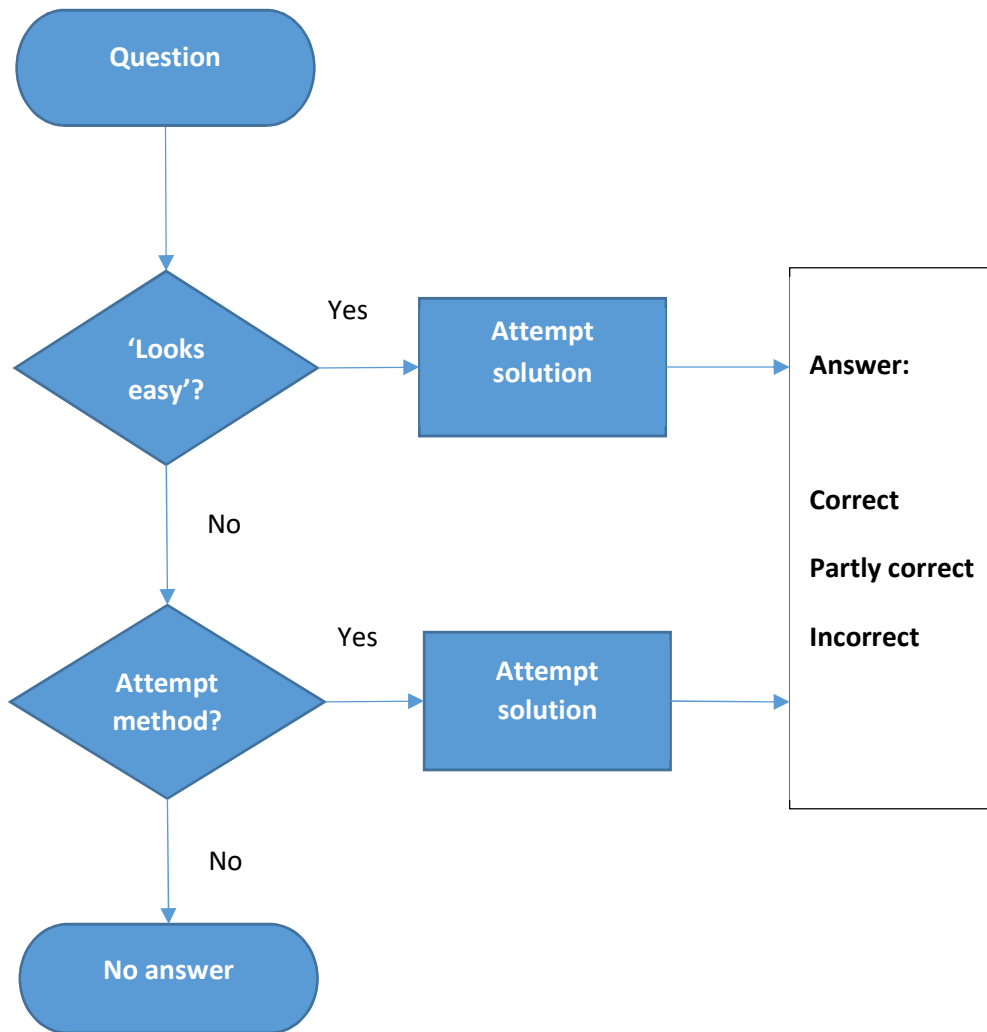
Among students in group Q, exactly the same tendencies occurred, even though these students had already submitted (incorrect) answers to the questions. They had had no feedback, however, so they did not know that they had been unsuccessful. Of question 2, students whose confidence was not matched by their ability wrote *'I understood what I had to do'* (student Q31F); *'easy to understand the question'* (Q32M); and *'fairly straight forward question'* (Q43F). Students with hasty and unproductive methods wrote, *'it was straight forward by only having red counters and simple steps'* (Q48F); *'use the numbers provided to easily add up and get the answer'* (Q49F); and *'it was simplistic addition. The overall probability must add to 1 so that made it much easier'* (Q136M). These patterns of answers were repeated across the other questions. There was no observable difference, then, between the approaches and justifications of students in the two groups who believed questions to be easy but in fact found them not to be so.

Applying to the looks easy theme the thinking of Marzano's and Kendall's New Taxonomy (2007), referred to in Chapter 3, alongside insights from cognitive science such as those by Willingham (2009) and those synthesised by Perry *et al.* (2021), an understanding can be constructed of the thinking processes of some of the students. When encountering a new question, a process of visual recognition occurs. Willingham describes the process:

'Thinking is *slow*. Your visual system instantly takes in a complex scene. When you enter a friend's backyard you don't think to yourself, "Hmmm, there's some green stuff. Probably grass, but it could be some other ground cover – and what's that rough brown object sticking up there? A fence, perhaps?" You take in the whole scene – lawn, fence, flowerbeds, gazebo – at a glance. Your thinking system does not instantly calculate the answer to a problem the way your visual system immediately takes in a visual scene' (2009, p. 6).

So it is for a student with an examination question. Before the student has even consciously engaged their thinking system, their visual recognition system has already made an instantaneous diagnosis: "it looks like probability: I know about this". The recognition ("looks like probability") may well be correct, although the assessment of competence (or self-concept) is perhaps more uncertain or, at least, as yet unproven. The student's 'self-system' then makes a decision about whether to engage or not with the question; a suggested flow chart for this process is illustrated in Figure 31, below.

Figure 31 - “Looks Easy” Theme: Judgement and Thinking Process



Source: Author’s own.

Given the context of a high-stakes public examination, where so much depends on the outcome, it is likely that only a few students will decide not to engage and, for some of them, this may be an anxiety-fuelled “fight-flight” response (see Rappleye and Komatsu, 2018; Putwain and Aveyard, 2018; Putwain and Symes, 2018). Assuming, then, that the self-system is engaged, the question “Looks easy?” becomes a real one, and the answer dictates a further decision. If the response is “Yes” and the student decides that the question looks easy then, following Marzano and Kendall (2007), they engage their metacognitive system, which

manages the thinking process that will attempt a solution. If the response to the “looks easy?” question is “No”, then a further decision takes place, governed by the student’s self-concept or self-efficacy. The student then needs to engage their cognitive system to decide whether they are willing to and/or can attempt a method. The student’s inner dialogue here might be, ‘Looks easy?’ ‘No, but I think I can manage it.’ This duality – not easy, but manageable – was visible in some student responses: student P10M wrote *‘haven’t done much revision but I know how it works’*; student P35M wrote, *‘we have covered this a lot in lessons and it takes a bit of time to workout though’*. In a similar vein, student Q02F wrote, *‘this one was a bit harder but still not too difficult. It took a minute to figure out the best way to go about the question’*. These students recognised the demands of the question, and they decided to engage with them.

If the answer to the second decision box question in Figure 32, “attempt method?” is “No”, then the student gives up and submits no answer. These two decisions may happen very fast indeed – there is a loop back to the self-system, as it were – and maths teachers are very familiar with students who declare themselves to be stuck before they have attempted any meaningful engagement with the problem (see, for example, Lee and Johnston-Wilder, 2018; Beveridge, 1997). It is outside the scope of this study to probe the reasons or mechanisms for this quick disengagement, but it is important to acknowledge the role that it plays in students’ responses to examination questions. It has already been shown in this study that this rapid disengagement – as evidenced by the lack of answers to many questions and the quick time taken to finish the questionnaire – led many students not to attempt questions to which they could probably have given at least partial solutions. In the context of a real GCSE maths examination, this split-second decision to disengage could make the difference between a lower and a higher grade, with all its consequences.

The looks easy theme is important because it represents part of a student’s initial reaction to an examination question. This theme develops comprehension of how students articulate their responses and, to an extent, their understanding of demand and difficulty in examination

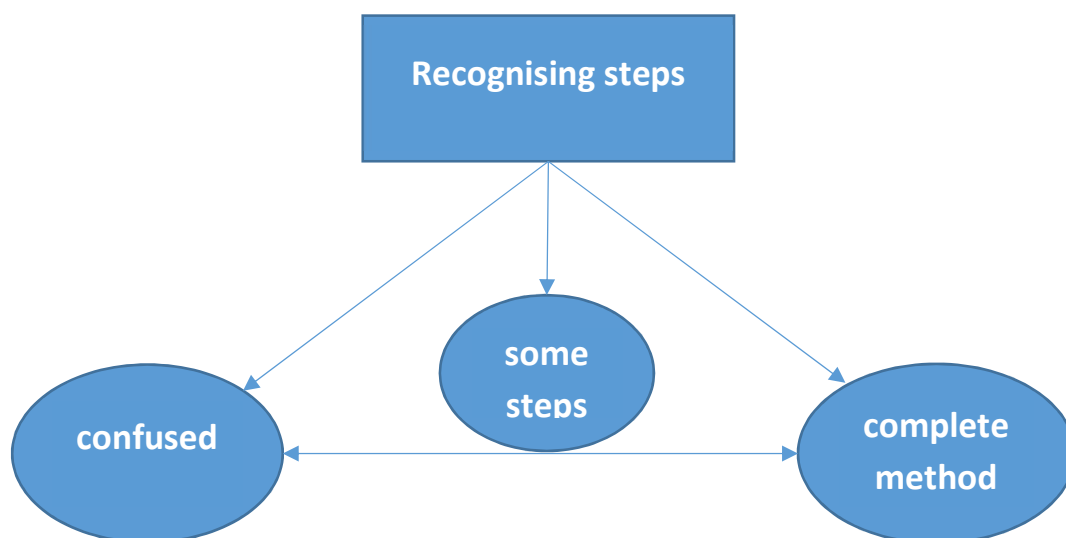
questions. Because deeper thinking is 'effortful, slow and uncertain' (Willingham, 2009, p. 6), account also needs to be taken of students' more considered and reasoned thoughts. These are under the second main theme, "recognising steps".

5.4.2 "Recognising Steps" theme

A student starts to engage with a question by seeking or "recognising steps" that will form a method. Responses around methods – complete, part, or entirely lacking – were common threads through the responses that students gave to explain or justify their estimates of question difficulty. There are three main ways in which students may recognise steps and engage with the demands of the question set by examiners. First, steps to answer a question may revolve around the recall of a formula or process that the students have previously learned. Secondly, some or all of the steps may be set out explicitly or implicitly in the question. Thirdly, there may be an absence of structure or scaffolding, and students may be required to construct a method for themselves. Students' responses to these various possible demands are varied, to an extent already noted, as in the free text answers quoted in previous sections of this chapter.

A number of graduated sub-themes (here noted within double inverted commas as coined terms) are contained within recognising steps. Some students will be "confused", and will not be able to remember, apply, or deduce any method; others will make "some steps" but not reach a complete solution; others again will be able to (describe and) apply a "complete method", perhaps involving several sequential steps. The process of recognising steps may, therefore, activate prior learning for many students. Prior learning (and understanding) is likely to have been reinforced by students' repeated practice.

Figure 32 - “Recognising Steps” Theme



Source: Author's own.

Figure 32 shows a representation of the recognising steps theme⁷¹. From the students' responses below, and from professional experience, we can see that when students recognise that there are steps that they need to take in order to 'crack' the demands of the question, they make their first approach to creating an answering strategy. Without this, the question will remain too difficult for them. Some students equated the number of different steps in a question with difficulty – the more steps, the more difficult the question (student P03F, for example, wrote *'lots of steps to do correctly which could take a while'*). This in contrast with pedagogical ideas of scaffolding, however, where steps are given that break down a question give access to students of all abilities.

⁷¹ To an extent, it is arbitrary as to how many sub-themes are recognised on the lower tier of the diagram, since there is a continuum between the extremes of offering a complete method and being completely confused. Nonetheless, it seems sensible to make distinctions between students who appear not to be able to make a start, those who have some idea but can construct only an incomplete method or solution, and those who are able to produce a complete method and a solution to the question that fully satisfies the examiners.

The low levels of student engagement in the pilot study are possibly explained by the fact that, in the context of a voluntary online questionnaire, sent out during lockdown, there was no incentive or external expectation for students to engage themselves with these questions. Many students appeared to have pushed through without spending too much time on their answers.⁷² This is in contrast to the main study, where engagement was at a higher level, perhaps because of the presence of the researcher/headteacher.⁷³

Students who described themselves as confused by the question, and stated that they “don’t know where to start”, were not able to identify any steps in a possible method, or were unable to apply a method stated or suggested in the question. These students scored no marks, except by guesswork and chance. The students may not have understood the point of the question, and their comments were irrelevant in relation to the demands of the question.

Other students recognised and were able to apply a few steps in a method. These students usually went on to enter at least a partial answer to the question, and gained some marks. If a student was able to recognise some steps, then it was possible that some more might subsequently follow, enabling the student to move more towards a fuller answer. A few students were able to articulate and/or apply a complete method. These students’ metacognition (control of their own thinking processes) was made visible through their descriptive answers. Confused students frequently declared that they were *‘not sure how to work out’* the question (P15F), that *‘it just looks confusing’* (P13F), or that *‘I didn’t know how to start it, leaving me confused’* (Q39M). Some students were able to go a bit further and explain the source of their perceived difficulty; explanations often focused on the presentation or wording of the question: *‘it was worded weird’* (Q64M); *‘layout/explanation is a bit complicated’* (Q62F); *‘I wouldn’t know where to start, because it has been worded quite*

⁷² The mean time taken for completion of the whole questionnaire was under 20 minutes.

⁷³ Students spent an hour completing the main study questionnaire in scheduled lesson time which, while they were free to give or withhold consent for their responses to be used, was spent under the direction of the researcher, who was also their headteacher. Engagement was consequently more prolonged and at a higher level in the main study.

confusing' (P14F); or *'there is trial and error which can be frustrating'* (P09M). These students focused on only one aspect of the question; their answers were relevant but vague, and they lacked depth.

Some students gave responses indicating that they recognised steps but they were not able to articulate a complete method. Within such responses, examples included, *'you had to work out a strategy first, and it wasn't obvious'* (student Q01N); *'just remember the method to do the equation'* (P21M – he did not specify which method, or which equation, so it is not possible to be sure about how much he understood); *'I think this would've involved a Venn diagram but wasn't sure how to answer it'* (Q23F); and *'add the probability's together to find the remaining one'* (P26F – not the correct method for solving question 2, but the student showed some idea about the arithmetic surrounding probability). More sophisticated, multi-step responses, included, for question 1, *'all you have to do is to take one number away from the other, then multiply it by 85, then divide by 100 for the cost'* (P24M – he had confidently summarised the first two steps of the method, and had converted the units, but omitted the final step of adding the standing charge). In a similar vein, student P33F wrote, *'just have to subtract miles and multiply with price'*.

Other responses took the form of an almost complete method, which might have made sense to the student at the time, but which did not communicate their understanding clearly enough; for example, in relation to question 5 (algebraic expressions), *'Parts A and B are basic equations, which I find fairly straightforward, and part C is quite easy because none of the values in the quadratic are negative'* (P47M). This last response is largely descriptive – the student has identified the essential features of an answering strategy, but he did not explain or present it adequately. Student Q40M wrote, for question 3 (conditional probability) that *'working out percentage of left handed males was difficult, but finding out how many weren't was easy'*. His answer was reflective in that it communicated the degrees of difficulty he

encountered in different parts of his answering strategy, but it gave no clear account of why, overall, he chose to allocate a particularly estimate of difficulty to the question.

A few students – less than 2% of responses overall, and exemplified by two students in group P – gave answers that showed an understanding of the different parts of the questions and the difficulty that each might bring. These responses, which tended to offer complete methods and explanations, were lengthy. In one instance, the student explained his choice of difficulty level for question 6, finding the circumference of a circle:

'All you need to do is double the radius to get the diameter, and then multiply this diameter by pi to get the answer (pi x d is the formula for the circumference). Therefore this question is easy, because you don't really need to do much working at all' (P47M).

Another student offered a pithier response: *'I understood diameter = 2 x radius, and circumference = pi x diameter'* (Q135M). He had included the right answering strategy but failed to link this to his judgement of how difficult he estimated the question to be, leaving an incompletely presented explanation. Student P48M, rather like P47M, similarly explained the rationale for his choice of difficulty estimate for question 5 (algebraic expressions):

'The question is neither easy nor difficult because there are three parts, the first parts being really easy and part c being fairly straightforward as long as you know how to factorize. Finding to [two] numbers that multiply to get 30 and that add to get 11. Then realizing that x can be one of those numbers' (P48M).

His response, of which this is the first part, shows how the different aspects of his question answering strategy had become integrated in a coherent whole. He recognised that some aspects of the question were more challenging than others, and he offered a strategy for solving the mathematical problem. His understanding of the topic was clearly more than adequate, but he did not go beyond this understanding to conceptualise or offer a higher level of abstraction.

A few students referred to their prior learning, and to repeated practice. This was particularly a feature of more developed and more confident student responses. Student P03F, for instance, reported that she had *'practiced algebra many times'*; P18M wrote that he had *'done it a lot so*

I'm good at it', and P30F suggested that she considered certain question to be easy *'because we have went over this multiple times in class'*. Sometimes the opposite was true, so that the lack of recent or repeated practice had undermined students' self-concept: *'I do not feel confident with probability as I haven't covered it for while'*, wrote student P38F. Bringing these two sides together, student Q39M wrote that *'I found the first two quiet [sic.] simple as I have been learning these in lesson, but I struggled on the last one due to me forgetting how to do the question'*; and student Q48F wrote, *'I've been learning how to do these recently so the method has stuck in my head but I'm not sure I know how to do [part] C'*.

Many students recognised the value of prior learning and repeated practice – these were mentioned by 40 students (31.5%), sometimes more than once; this recognition was summed up by a comment from student P39F, who wrote,

'I am confident with this topic and therefore find the questions easy to answer. I personally think it is because we do sort of 'making it stick' practice on small questions like these because it is almost certain that questions like these will come up on the tests.'

This student also wrote about how, in her maths lessons, *'every year we touch back upon the basics and add more relevant knowledge towards [our understanding]'* (P39F). She recognised how the incremental acquisition of skills in mathematics over time had helped develop her understanding. Experience such as this enabled this student, and many others represented in this study, to recognise steps that led to the creation of suitable methods to approach and solve mathematical examination questions. Prior learning and repeated practice will be discussed in more depth in the main study.

5.5 Evaluation of Pilot Study – informing the Main Study

The pilot study was carried out under post-COVID restrictions that were not ideal conditions for collecting rich data from students. However, there were some positive lessons learned from the study, which shaped the design and method of data collection in the main study that follows. Evaluative comments and implications for the main study are collated in Table 16 below.

Table 16 - *Evaluation of Pilot Study and Implications for Main Study*

Pilot study evaluation	Implications for main study design and method
Survey items	
Students needed to attempt the examination questions first , in order to give better quality responses about their experience of difficulty, as well as estimates of expected question difficulty that more closely corresponded to their actual experiences of question difficulty	All students in the main study will attempt the mathematics questions before answering the survey
The six past paper GCSE mathematics questions enabled a wide range of topics to be included in the study, broadening its scope	I intend to use the same mathematics questions in the main study, to keep the range broad and to ensure that direct comparisons can be made, if useful
Sample size and composition	
Although 127 students were sampled in total, these came from 5 different schools . Only two schools had more than 30 students participating, and 2 had less than 20 students. This made it hard to tell if variations between respondents were associated with differences in teaching, or any other factors	It would be better to recruit students from a similar background , to minimise the effect of factors such as possible variations in teaching or curriculum coverage
Although the sample size was reasonably large (127), students were able not to enter responses for questions if they wished; for one question, the number answering was only just over half of the total. This may have skewed some of the comparisons (e.g. if a student did not answer, should it be assumed that they found the question difficult?)	Encourage all students to give responses for all questions , and make sure they have enough time to do this

Pilot study evaluation	Implications for main study design and method
Student engagement and quality of response	
Student engagement with the survey instrument was fuller if they had some connection with the researcher : the fullest responses came from students at my own school	I shall locate the main study entirely in my own school , and supervise all the data collection personally
Students need to take time to create rich responses that tell the story of their experience and comprehension of difficulty in examination questions	I shall allow sufficient time for the data collection – 1 hour per class
It would be valuable to explore some of the students' responses with them more fully , in a less structured and more dynamic way. This was not possible within the constraints of the pilot study and COVID restrictions. It could possibly have been done via video conferencing software, which was not tried, but it experience suggests it is difficult to achieve a satisfactory group dynamic through this medium	Following the collation of survey responses, I shall recruit students for in-person focus group discussions , to explore their ideas, experiences and understanding more fully
5-point Likert-type scale appeared to give sufficient flexibility for a range of student estimates of difficulty. A central tendency was evident in some responses	No obvious reason to change the style or number of response options. Increasing the number of options would be unlikely to affect the central tendency
Opportunities to extend the scope of questioning	
Going beyond students' estimates of question difficulty and explanations of their answering strategies, it would be valuable to learn more about their views of comparative difficulty , and the factors they attribute to this	I shall include survey items asking which question students found most/least difficult , and asking for reasons
There were no opportunities given for students to compare questions on the same topic , which might have enabled them to think more evaluatively about the factors that made one more (or less) difficult than the other	I can introduce an additional mathematical question on one of the topics covered in the pilot study, to enable this comparison to be made
There could have been an opportunity for students to reflect more generally on factors that contribute to demand and difficulty in examination questions	I will include an additional survey item in the main study, asking this more general question

Source: Author's own

5.6 Conclusion of the Pilot Study

In this chapter, the results of the pilot study in which 127 students from 5 different secondary schools considered 6 GCSE mathematics past paper questions were reported and analysed.

From their estimations of difficulty, their explanations of their choices, and the marks they gained for their answers, students' variable understanding of how they experienced demand and difficulty in GCSE mathematics questions was revealed. At times, students were confused by a question's requirements, whether through the inherent demands of the topics that were presented, or through presentational aspects of the question, some of which may have presented them with difficulties not foreseen or intended by examiners, as in Question 2.

Some students were able to recognise at least some of the steps that were required of them to construct an answer to a given question. This study has demonstrated that students' answers and explanations, at their most fully developed, can give a clear understanding of the aspects of demand and difficulty operating in GCSE examination questions, and that students who are aware of these aspects are able to exercise a great deal of control over their own metacognitive processes. On the other hand, this study has also shown that, for many students, judgements about whether a question "looks easy" – which may be made unconsciously and without engaging the cognitive system – can lead them too quickly to make assumptions that do not match the demands of the actual question asked.

Searching across the whole pilot study data set of more than 2,000 individual responses and answers, repeated patterns of meaning in the data have been reported upon and discussed within this chapter. Through a long process of reflexive thematic analysis, themes and sub-themes have also been identified and exemplified; each has been both grounded in the students' own words, and has been related to aspects of learning theory, including Marzano's and Kendall's New Taxonomy (2007).

Although the sample size was adequate (127) for a qualitative pilot study of this nature, at times the engagement of the students appeared low, and the completion rate for all but one

of the mathematical questions was below 80%. In order to provide a larger sample and a fuller set of student responses for analysis, it would be good to take steps to improve a number of aspects of the study, ready for the next stage. Some valuable lessons were learned, and these will now be applied to the main study that follows.

This page has been left intentionally blank

Chapter 6: Main Study – Exploring the new world in more detail

In this chapter I have rich and immersive encounters with residents and fellow travellers on my journey, in the form of the main study and my reflexive thematic analysis of my encounters.

The main study involved 97 students from four classes in Year 11 at one comprehensive secondary school in Gateshead, North East England, between 21 and 27 September 2022. 95 students answered a sample of past paper GCSE Mathematics questions and completed a questionnaire. Eight of these students also took part in two focus group discussions, on 19 and 20 December 2022, along with two additional students who had not been present when the questionnaires were completed. It was found in the pilot study that students gave more coherent explanations of their views about examination questions if they had attempted the questions (and not merely looked at them) before discussing them. In the main study, therefore, all students were asked to attempt the mathematics examination questions before answering the questionnaire. In the first part of the questionnaire, students were asked to rate the difficulty of the mathematics questions they had just attempted, using a 5-point Likert-type scale with responses ranging from “very easy” to “very difficult”. The second part of the questionnaire asked for free text responses, probing students’ views on the relative difficulty of some of the questions they had answered and the reasons for the estimations of difficulty. All students completed their answers on paper, in the presence of the researcher, who was also their headteacher. No electronic alternative was used. Levels of engagement were high, and all students completed all parts of the question paper and questionnaire. Copies of the questions and the questionnaire appear as Appendices C and D.

Students’ answers to the mathematics questions were marked, according to the examination board’s mark scheme. These marks and students’ responses to the first part of the questionnaire (estimates of difficulty) were analysed, using straightforward statistical methods to find frequencies of responses, the mean and mode average, and the standard deviation.

Correlations between student's estimates of difficulty and the marks they obtained were also analysed, using Pearson's correlation coefficient (Pearson's r). These results are collated and presented below. Students' written responses to the second part of the questionnaire, and their focus group remarks, were coded and analysed using thematic analysis techniques, as described by Braun and Clarke (2022). The results of these analyses are reported below, in sections 6.2 to 6.6. Anonymous codes were attached to student questionnaires before analysis. These codes take the form of the letter 'R' (which is the code for the main study, following codes 'P' and 'Q' used in the pilot study), followed by a 2-digit sequential number, followed by a suffix letter indicating the preferred gender of the student (M = male; F = female; N = non-binary or prefers not to say). A student might be identified, for example, as R17F, R36M or R15N. These same codes were used for the students who volunteered for the focus groups.

6.1 Data analysis: descriptive and inferential statistics

6.1.1 Sample size and gender distribution

Students were drawn from four mathematics classes at one comprehensive secondary school. Students are streamed by prior attainment for mathematics at this school. The classes were all from the top and middle streams (or bands). All 95 students who attended their lesson on the day the survey was carried out participated in the study: they were given the option to opt out, but none did, either at the time or subsequently. There were slightly more female than male students in the sample. In the school in which the study was carried out, there were slightly more female than male students in the Year 11 year group; there were more female than male students in the top and middle bands for mathematics, and more male than female students in the lower band for mathematics.

6.1.2 Estimates of difficulty

Table 17 - Main Study, Students' Estimates of Difficulty

Students' estimates	Q1 Algebraic expressions	Q2 Prob-ability	Q3 Conditional probability	Q4 Growth and decay	Q5 Algebraic expressions	Q6 Mensuration	Q7 Probability	Total %
1 Very easy	32	12	4	28	56	28	73	35.2%
2 Easy	48	35	13	27	22	19	15	27.1%
3 Neither easy nor difficult	11	29	27	24	14	21	5	19.8%
4 Difficult	4	16	37	14	2	18	1	13.9%
5 Very difficult	0	3	14	2	1	6	0	3.9%
Total	95	95	95	95	95	92	94	
Mean	1.9	2.6	3.5	2.3	1.6	2.5	1.3	
Mode	2	2	4	1	1	1	1	
Standard deviation	0.8	1.0	1.0	1.1	0.9	1.3	0.6	
Rank by mean (most difficult first)	5	2	1	4	6	3	7	

Source: Author's own

In this study, as shown in Table 17, students tended to find the questions easy or very easy, with the exception of question 3: "very easy" was the most popular choice overall. The middle option, "neither easy nor difficult", was popular in several of the questions, but it was not the most popular choice for any individual question. Perhaps because the students had already answered the questions, their estimates of difficulty avoided a central tendency, as was also observed in group Q of the pilot study.

6.1.3 Students' marks for GCSE Mathematics questions

Marks that the students gained in their actual answers to the mathematical questions, and their associated statistical measures, are reported in Table 18 (below). The number of marks available per question varied: most questions had 4 marks available, but question 3 was out of 5 marks, and question 5 was out of 6 marks. As well as the marks that students gained, the proportion who attempted each question is reported, along with the proportion of those who attempted each question, and those who gained either no marks or full marks.

Table 18 - Main Study, Students' Marks

Marks and Attempts	Q1 Algebraic expressions	Q2 Probability	Q3 Conditional probability	Q4 Growth and decay	Q5 Algebraic expressions	Q6 Mensuration	Q7 Probability	
Marks:	0	3	33	46	15	0	54	1
	1	3	14	10	12	1	0	5
	2	12	4	0	7	0	5	13
	3	3	3	39	24	13	36	76
	4	74	41	0	37	3		
	5		0		29			
	6				49			
Attempts	95	95	95	95	95	95	95	95
% attempted	100%	100%	100%	100%	100%	100%	100%	100%
Fully correct	74	41	0	37	49	36	76	
% attempts fully correct	77.9%	43.2%	0.0%	38.9%	51.6%	37.9%	80.0%	
% attempts gaining zero marks	3.2%	34.7%	48.4%	15.8%	0.0%	56.8%	1.1%	
Mean mark	3.5	2.1	1.3	2.6	5.2	1.2	2.7	
Mode mark	4	4	0	4	6	0	3	
Standard deviation	1.0	1.8	1.4	1.5	1.1	1.4	0.6	

Source: Author's own

All students in this study attempted all seven questions. Some of these questions did not appear to discriminate particularly well between students: most students gained full marks in questions 1, 5 and 7, and a small majority gained zero marks in question 3.

Table 19 - Most and Least Difficult Questions

Most and Least Difficult Questions	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Total
Least difficult – frequency (students)	6	1	2	1	26	6	39	81
Most difficult – frequency (students)	0	11	55	7	1	13	0	87

Source: Author's own

Students were asked to say which questions they found most and least difficult. They were also asked to compare the difficulty of questions 2 and 7, which were both on the topic of probability. The distribution of their answers is given in Table 19, above. Some students entered more than one answer, so their responses were disregarded.

Students were split in their views about the least difficult question between Question 5 (32.1%) and Question 7 (48.1%). Most students (63.2%) thought that Question 3 was most difficult.

The reasons given by students for choosing a question as the least/most difficult did not appear to depend on the actual question they chose. Their responses for 'most difficult' often mirrored or complemented their responses for 'least difficult.' For example, student R86F wrote that Question 5 was least difficult because 'I was familiar with the type of question and knew how to work it out with the correct method,' whereas Question 3 was most difficult because 'I didn't know how to do the question or even how to start and didn't know a method.' Student R87M chose Question 7 as least difficult because 'Probability is easy as we have had a lot lessons on it,' and he chose Question 3 as most difficult because 'I couldn't remember exactly how to do it.' These responses have been coded yellow for methods and green for

memory/familiarity/practice. Students' reasons for identifying questions as least and most difficult are summarised in Table 20 below.

In terms of thematic analysis, this process is concerned with generation of initial codes, corresponding to step 2 of Braun's and Clarke's 6 steps. Because the reasons for 'most difficult' often mirrored those given for 'least difficult' questions, these have been set out in (roughly) opposite pairs, where possible. The detailed responses students gave for these questions were included in the thematic analysis for the whole survey, and they will be discussed later in this chapter.

Table 20 - *Main Study Questionnaire: Reasons Given for Least/Most Difficult Question*

Reasons given: least difficult	Freq.	Reasons given: most difficult	Freq.
Straightforward arithmetic	30	Struggled with maths and method	20
Simple presentation, easy to understand wording	7	Unclear presentation, hard to understand wording	23
Remembered how	7	Forgot how	17
Recent revision or practice	15		
		Hard topic	3
		Uncertainty or confusion	13

Note: not every student gave an answer as to why they had identified a question as least/most difficult.

Source: Author's own

In their views of the relative difficulty of two probability questions, students were almost unanimous: 94.6% of them decided that Question 2 was more difficult than Question 7. The text of these questions is printed here, for ease of comparison.

Question 2.

Finn has two bags of counters.

He takes a counter at random from each bag.

The probability that he takes a red counter from the first bag is 0.3

The probability that he takes a red counter from the second bag is 0.4

What is the probability that he takes at least one red counter? (4 marks)

Question 7.

A bag contains 12 counters.

6 are red, 4 are blue and 2 are yellow.

A counter is taken from the bag at random.

What is the probability that the counter is a) Red b) Yellow c) Green (3 marks)

The relative demands of the mathematics questions in the student questionnaire, using the CRAS scales of demands, have already been given in the report of the pilot study (section 5.3). With the addition of Question 7 in the main study, the demands of this new question are also evaluated, using the CRAS scales, and the comparison is given in Table 21 below. Question 7 was chosen because it was similar in style to Question 2 – both questions involved taking a coloured counter out of a bag, which is a familiar context for probability questions in the classroom. The demands are very different, however: in Question 7 the response strategy is extremely straightforward, and because there is only one event, there is no need to use a probability tree or multiply fractions. Question 7 was arguably *too* easy. Students' responses showed they agreed: R01M was blunt – '*primary school fractions, like very simple.*'

Table 21 - *CRAS Scales of Demands for Survey Questions 2 and 7*

Question	Complexity	Resources	Abstractness	Task Strategy	Response Strategy
2. Probability	3: more complex and inter-dependent ideas	2: most of the information needed is given	4: mostly abstract	4: students need to devise their own strategy and monitor its application	1: organisation of response is very straight-forward
7. Probability	2: single ideas and simple steps	1: all (and only) the information needed is given	3: abstract, but related to simple and concrete constructs	2: strategy is inferred	2: organisation of response is straight-forward

Source: Author's own

The only slight complication was the inclusion of part c) which asked for the probability of choosing a ‘green counter’ – when there were no green counters. Few students fell for that. R25M was typical: *‘I read the question and didn’t fall for the trick question.’* One of the four students who identified Question7 as being *more* difficult, stated that *‘I found the question itself to be easier, however it took longer to realise the simplicity of the question due to having overcomplicated it at first and looking for methods that would have been unnecessary’* (R03F).

Evaluating the inclusion of this question in the study, it is possible to explain that it was chosen as a comparison because the available probability questions from past papers using the “counters out of a bag” context were limited, but in this study it did not provide much of a challenge, and it was therefore a poor comparison. It was not possible to learn much from students’ responses, except to observe that students found it particularly straightforward to compare the relative difficulties of the two questions.

6.1.4 Correlations

The strength and direction of the linear correlation between students’ marks and their estimates of difficulty was investigated, using Pearson’s correlation coefficient (r). This data is report in Table 22 below. As reported in Table 22, among the students in the main study and for the majority of the questions, the marks they gained and their estimates of difficulty were moderately negatively correlated.

Table 22 - *Correlations between Students’ Marks and their Estimates of Difficulty*

Measure	Q1	Q2	Q3	Q4	Q5	Q6	Q7
Correlation coefficient (r)	-.32	-.31	-.48	-.57	-.40	-.38	-.10
Sample size (n)	95	95	95	95	95	95	95
Significance (p)	.001	.003	<.001	<.001	<.001	<.001	.328

Source: Author’s own

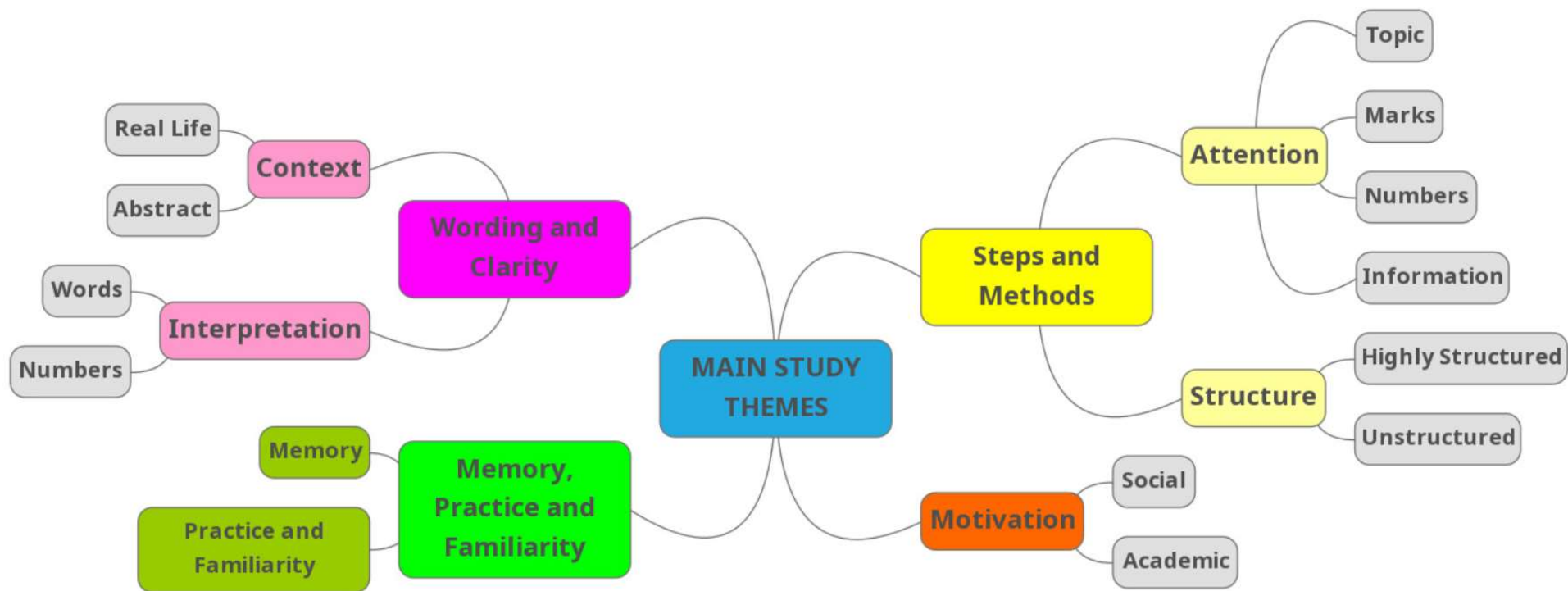
The exception to this was Question 7, where there was no evident link. The strongest negative association between marks gained and estimated difficulty was observed in Question 4, $r(93) = -.57, p < .001$. (This association was also strongest for Question 4 in the pilot study.) Since students' actual marks have an inverse relation to their experienced difficulty of a question, it is possible to state – with the exception of question 7 – that their estimates of difficulty were moderately correlated with their experienced difficulty. Analysis of the answers and difficulty estimates of male and female students showed no significant differences.

6.2 Reflexive thematic analysis

Following a process of reflexive thematic analysis, as described in Chapter 4, four main themes were developed. These themes are Steps and Methods; Wording; Practice, Memory and Familiarity; and Motivation. A diagram showing the mapping of these themes appears as Figure 33 (below).

These themes are now explored in turn, illustrated with suitable examples from students' responses. Students' responses in this study were full and rich. Students' responses in the focus groups were particularly well developed, and these were transcribed verbatim, including hesitations and fillers, to present in this thesis the authentic voice of students. Student responses have therefore been quoted at some length, to capture the full range, context and meaning of their thoughts.

Figure 33 - Mapping Themes from the Main Study

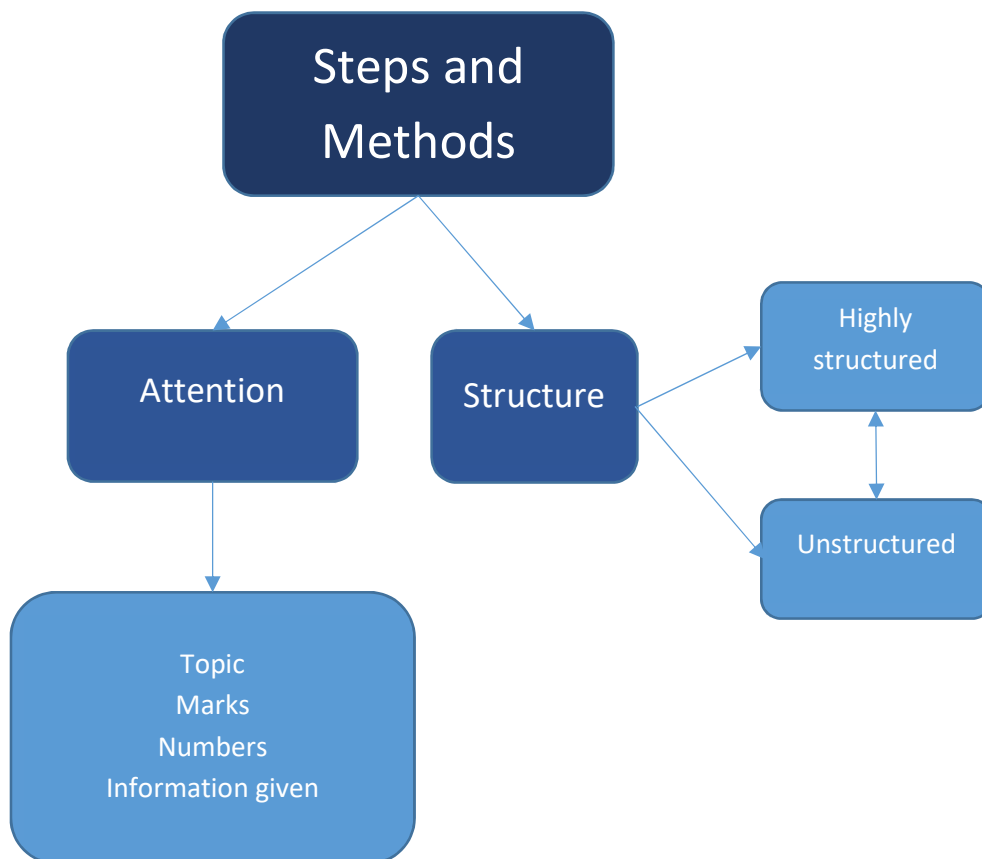


Source: Author's own

6.3 “Steps and Methods” theme

Steps and methods were frequently cited by students in this study as adding to the difficulty of questions (30 students commented: 17 males and 13 females), and also as factors that made questions easier (12 students commented: 7 males, 5 females). Students’ comments were developed into two main sub-themes, exploring where students focused their **attention** when they encountered a question, and the **structures** – or lack of them – that they brought to their responses. This theme can be illustrated diagrammatically, as in Figure 34 below.

Figure 34 - “Steps and Methods” Theme



Source: Author’s own

Several students commented in a general way on the relationship between steps and methods and the difficulty they experienced. One stated simply that *'questions with multiple steps can be difficult'* (R09M). Contrasting two questions, another student explained this relationship straightforwardly. Commenting on question 3, which she had found more difficult, she said that,

'[I] don't know what method would be used to work it out, and I also had a harder time figuring out what steps would have been required to find the answer' (R03F).

Whereas in question 7, she explained that,

'I found the question itself to be easier, however it took longer to realise the simplicity of the question due to having overcomplicated it at first and looking for methods that would have been unnecessary' (R03F).

It is interesting to see this student reflecting that she overthought the question at first, perhaps expecting it to be more difficult than it subsequently turned out to be. In general, students equated the need for more steps within a method with increased question difficulty. It is important to delve deeper, to find out why this might be. Asked to compare the difficulty of questions 2 and 7 on the questionnaire, which tested the same topic (probability), students were able to identify a number of differences:

[Q2 was more difficult] *'as it is more steps to get the answer and not as straight forward as question 7'* (R55F).

[Q2 was more difficult than Q7] *'This is because there was more steps and numbers to think about than the simplicity of putting the numbers into a fraction'* (R39F).

[Q2 was more difficult than Q7] *'This was because firstly it had more steps to get to the answer and involved some working out such as multiplying fractions whereas question 7 I could just look at it and I could easily answer the question'* (R50M).

'Question 2 consists of more steps and calculations. This makes it harder because it results in the student thinking harder about working more' (R49M).

There is both a quantitative and a qualitative element to these students' graduated responses. Student R55F was typical of several students in asserting plainly that more steps create more difficulty. Student R36F added that more steps were accompanied by more numbers, perhaps implying an increased cognitive load caused by the requirement to select and appropriately

manipulate the numbers. Student R50M was typical of some other students in going one step further. As well as referring to the addition of more steps, he explained that question 2 demanded a more complex operation – multiplying fractions – which made the question more difficult than (it may be inferred) straightforward arithmetic that he *'could just look at'* and *'easily answer the question.'* Student R49M went further again in his analysis: he was aware that question 2 required him to think harder, and it was this effortful thinking that was the measure of difficulty.

Several students felt that the addition of more steps required to answer a question introduced a greater possibility of error. Although the following students identified different questions on the questionnaire as being more difficult, their reasons for determining that difficulty were similar:

'You have to do a lot of steps and it can be easy to make a mistake' (R30M).

'It was a multiple step process which makes mistakes more likely' (R34M).

'Multiple step questions [are more difficult] because you can carry over mistakes and simple step questions you may be forced to reconsider and so spot mistakes' (R37M).

'When more steps are required to get to the answer or it takes more time and increases the chance to mess up ruining the whole answer' (R25M).

For a number of students, however, it was not only the number of steps involved that contributed to question difficulty, but the nature of the procedures involved in these steps, and perhaps also the inherent complexity of particular topics. There was no consensus on which mathematical procedures were perceived to be most difficult: students mentioned algebra (3 students), multiplying fractions (3 students), having to complete the square (1 student) and interpreting diagrams (1 student). In contrast, 6 students mentioned carrying out simple arithmetic and simplifying fractions as procedures they found less difficult. One student claimed *'I find it easy to factorise simple algebra and quadratics because there aren't a lot of steps involved'* (R24M). No other students mentioned factorisation, algebra and quadratics as

being less difficult topics – quite the opposite, in fact, with algebra cited by 3 students as a feature that *added* to the difficulty – but this student’s explanation of the small number of steps involved did resonate with other responses.

Turning to the responses that students gave in the focus groups, thematic analysis of these more in-depth answers and discussion led to the development of the sub-themes of attention and structure.

6.3.1 Attention

One of the questions explored with students in the focus groups in this part of the study was where they focused their attention when presented with an examination question in mathematics: what did they look at first? Members of the focus groups gave a variety of answers, showing some sophisticated approaches to examination questions as well as a degree of confidence. Some of them scanned the question but focused more on the end of it, looking for instructions and important information, and for the mark allocation:

[Researcher:] When you look at an examination question, what do you look at first?

‘Usually the marks, like how much it’s worth, and you see how long you can spend on it’ (R34M).

‘I’ll look at the marks as well, cos it, like, tells us how much work I need to do, cos there’s a difference, obviously, between like a 2-mark question and a 5-mark question. So straightaway I can see if I need to work out more to do with the numbers, or if I’ve missed out a step and it’s more marks than, like, a 2-marker’ (R36M). [...]

[Researcher:] Would you always use the same approach, or does that vary from question to question?

‘If it’s like a longer question and multiple lines and, like names and things like that, I would look at the marks first, cos that would tell us; but if it’s a short, like, “work out”, like two fractions for example, then I could just see what I need to do, I could just use the numbers straight away, without looking at the marks, cos I would just, I would know what to do straight away’ (R36M).

These students, then, are used to evaluating the difficulty of the question by having a quick look at the number of marks before they even get to grips with the question itself. It is evident,

from these comments that relatively expert students⁷⁴ are making quick but conscious judgements about the amount of time and effort to devote to each question, based on the complexity of the command verbs and the mark allocation. This contrasts with the approach of other, less confident students in this study, who were less sure of where to focus their attention: for example,

'It's hard to think of what to do as you don't know all the numbers and it's hard to figure out methods to find the missing information' (R17F).

Students in the focus groups referred to their frequent practice in lessons, indicating that they are well drilled in recognising and responding to different question types. The use of terms such as a "2-marker" in student R36M's answer points to his familiarity with the different formats and demands of examination questions.

Another student said that he would scan the question for its instructions:

'I tend to look for just the very end of the question, just exactly what it's asking for... and then I can just find the important bits of information in the question, like the numbers and that, instead of reading the whole question' (R31M).

A few other students – all males – also referenced having the confidence and expertise to make a very quick judgement: for example,

'I just look at the numbers and go, "ok, that's what I need to do"' (R77M).

'I usually just look at the numbers first, to be honest, and sort of figure out what I need to do from there' (R49M).

Other students, however, were more methodical in their search for the essence of the question:

'I usually, like, look through all the numbers first, and I'll see what the question's actually asking, and then I'll go back to the numbers and then work out important, like, differences between them and everything like that' (R51M).

⁷⁴ Students in the two focus groups were from the top and middle bands of a streamed system in a comprehensive school, and all were expected to gain a grade 6 or above in GCSE Mathematics.

'First of all I look for what it's asking us to do. And then I look at what I've actually been given in the question, like what numbers I've been given. Then I just go from there really' (R96F).

This is not yet a description of a methodical approach, but it shows the students scanning the question for important numbers and information, from which they will then begin to formulate an approach. As will be seen in the following section, this approach could be more or less structured. Another student explained his focus differently, although also in a search for how to tackle the question:

'The first thing I'd focus on in a maths question is, like, what kind of topics it's bringing in, because I can distinguish what I actually need to do to get, not the entire answer but at least some marks' (R77M). [...]

[Researcher:] So why is the topic important?

'Because each, er, each topic has different ways to go about things, and if you can work out what topic it is, you can work out the general formation of how to get the answer' (R77M).

This student had learned that particular topics are associated with specific methods to solve problems. In his initial analysis of the question, therefore, he is consciously looking to identify the topic, so that he will be able to bring a suitable method to bear on the specific demands of the question. There is an element of metacognitive understanding evident in this student's response.

When asked if their approach would differ in an examination from their approach in class, all the students in the first focus group said that they would take a similar approach, although one added that *'If I'm, like, in a test, I'd show my working out, whereas in class I'd usually do it mentally in my head'* (R51M). This student would take care to show his working out in an examination situation, which is often allocated marks in the examiners' mark scheme.

In summary, students in the focus groups described their initial approaches to examination questions in terms of their search for understanding of the requirements of the question.

Some described looking for the topic of the question first, which might lead to the selection of a suitable method. Others described how they look at the number of marks allocated, to gain

an initial idea of the likely complexity of the question. Most said they would also appraise the numbers involved, along with the other information in the question. After these initial scans of the question to evaluate its demands, students described how they would proceed to the application of a mathematical method.

6.3.2 Structure

Students' initial approaches to the question provided a range of responses around a similar theme, showing that they were more methodical – or less methodical – in their thinking. When they came to discuss the construction and application of methods in addressing the demands of the examination question, students' approaches were similarly more structured, or less structured. Figure 35 shows just two opposing points – 'highly structured' and 'unstructured' – but this is in effect a continuum, as implied by the double-headed arrow between these, and many midway points ('slightly structured', 'rather unstructured', etc.) could be identified. On the unstructured end of the continuum, some students described a rather haphazard approach:

'First of all I look for what it's asking us to do. And then I look at what I've actually been given in the question, like what numbers I've been given. Then I just go from there really' (R96F).

[Researcher:] So you're looking for what it asks you to do and what numbers you've been given, and then...

'... I dinnaa... I just see if I can try different ways of what could work. And if I end up getting something where I'm like, I'm thinking I'm along the right track, I just keep going in that direction' (R96F).

This student has adopted a "trial and error" approach, with only a superficial level of analysis of the question demands. Then she sets off on her method, in an unstructured way. There is, however, also a superficial level of control and evaluation implied in her response: if her approach appears to be working, she keeps going. Critically, though, this student is not asking herself any questions about whether the method she is applying is the most appropriate to the

question's topics or demands, nor whether it is the most suitable or efficient method she could use. It would be possible to describe this approach as 'incompetent', in that it is not intentional, evaluative or informed by her prior experience. But the student is at least aware of it: she is therefore 'consciously incompetent' in her approach to questions. Her approach may well work, at least for less demanding questions, unfortunately reinforcing her strategy, but it is likely to fall down when it comes to more complex problems.

Most students adopted approaches that were more structured. Referring to question 1 on the questionnaire⁷⁵, the majority of students took a similar and highly structured approach. This question falls into the category of algebraic expressions on the examination board's specification, but students did not need to use algebra to solve it. The following succinct student explanation was typical of many in the focus groups:

'What I first did was find the difference between the different mileages and then I times'd it by 0.85 to find out what's the extra mileage he had, and then added a hundred and fifty pounds to that' (R77M).

Question 1 was described as very easy or easy by 84% of students in this study, and 78% of them gained full marks. It is not surprising that many were able to articulate structured and effective methods to solve the question, therefore.

Question 3⁷⁶, in contrast, was one of the most difficult questions on the questionnaire. In this study, 54% of students described this conditional probability question as very difficult or difficult; no student gained full marks, and 48.4% gained no marks at all. Focus group students outlined a number of different approaches and methods, including ratio, tree diagrams and a table. In the discussion, these students had access to the question, but they did not have their own answers (from three months earlier) in front of them, so they were speaking from

⁷⁵ Question 1: Reuben hires a car. It costs £150, plus 85p for each mile he travels. When Reuben hires the car, its mileage is 27,612 miles. When Reuben returns the car, its mileage is 28,361 miles. How much did Reuben pay to hire the car? (4 marks)

⁷⁶ Question 3: 60% of the people in a town are males. 20% of the males are left-handed. 21.6% of all the people are left-handed. Work out the percentage of the people who are not male who are left-handed. (5 marks)

memory. It is worth quoting the students' responses at length, to capture the detail of their approaches:

'I don't know if I got it right, to be honest... I tried to do some weird thing – like, I don't know if I did a ratio or something, where I did, like, the number of men, the number of left-handed and tried to work it out from there' (R34M).

'To begin with, I'd times the percentages by a hundred and then put them into a tree diagram, so I could, like, see what was being said. So, for example, I'd put 600, like, on top and then a line being drawn down to females and then on the line to males I'd put like 400 right-handed males and 200 left-handed males – and then continue like that' (R51M).

'I think I put it into three separate tables of males, females and the whole population, and then have a right-handed and a left-handed, and then try to work out a value for each place in the table. And then, for the question "who are not male who are left-handed" I think I looked at how many overall were left-handed and then saw, and then just took my value from the table which is easy enough to find, to work out if you've got the three values that you're given, and then you could just, like, insert them into a fraction and then turn the fraction into a percentage to get your answer' (R36M).

[a page break has been inserted here, to enable diagrams and discussion to appear together]

It is interesting to see the varying levels of structure within these three students' responses. It would be perfectly possible to use a ratio method to solve the problem, but student R34M, who suggested this method, was not clear about how he would apply it in this case. Nonetheless, it can be seen from his answer sheet below (Figure 35) that he was able to use this method to reach a response that was partially – but not fully – correct.

Figure 35 - Ratio Method from Student R34M

3. 60% of the people in a town are males.
 20% of the males are left-handed.
 21.6% of all the people are left-handed.

R34M

Work out the percentage of the people who are not male who are left-handed.

Town

m : N M
 6 : 4

40 % of town not male.

males

Left • Right
 hand • hand
 2 : 8

All people

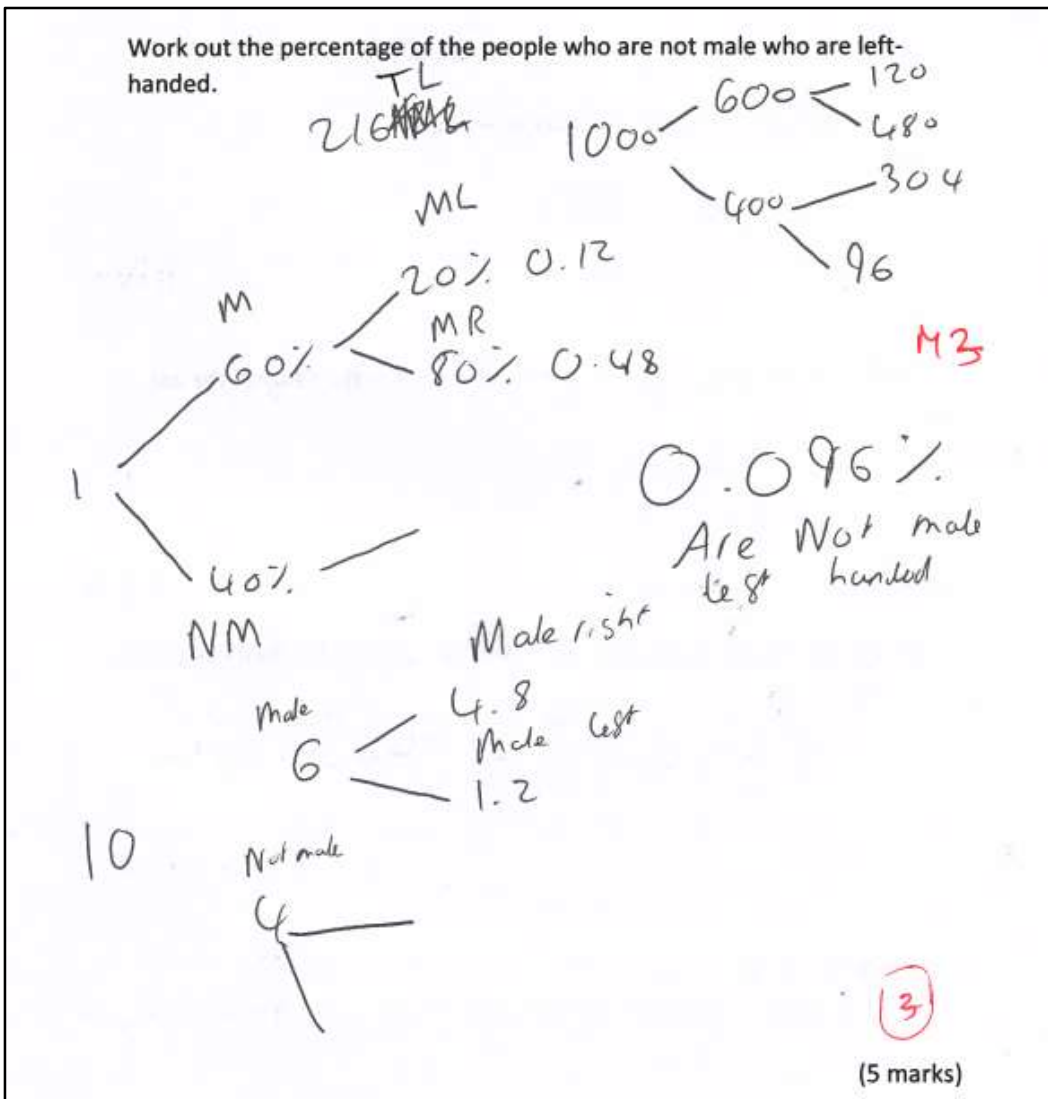
Left • Right
 hand • hand
 2.16 : 7.84

12 % of whole population is left handed male.
 21.6 - 12 = 9.6

Source: Author's own

Student R51M, who suggested a tree diagram, was clearer about how his method would work. His answer paper is shown in Figure 36, below. It is not necessary to convert the percentages into numbers for this method to work, but possibly this student preferred to visualise and work with whole numbers rather than percentages.

Figure 36 - Tree Diagram from Student R51M



Source: Author's own

The fullest description of a method came from student R36M, with his use of tables. His answer sheet (see Figure 37, below) shows a rather loose tabular approach. Rather than using separate mini-tables, it might have been clearer to use separate rows within the same table for males, females and the whole population, and columns for right-handed and left-handed, for example – but his structured approach was partially successful, in that it allowed him to reach a response that showed the proportion of the *total* population that was both not male and left-handed.

Figure 37 - Table Method from Student R36M

Work out the percentage of the people who are not male who are left-handed.

~~60%~~ ~~40%~~

60% 40%

12% → M+L 21.6 - 12 = 9.6%

<p>600M</p> <p>120L</p> <p>480R</p>	<p>400F</p> $\begin{array}{r} 1216 \\ - 120 \\ \hline 096 \\ 96 \\ \hline 1000 \end{array} = 9.6\%$
-------------------------------------	---

Source: Author's own

Another student offered an answer that demonstrated an element of reflection and metacognition resulting from his memory of a less structured approach:

'I didn't do it as a tree diagram, but I feel like that would have made a lot more sense than what I did. I think I tried to do, for instance, 60% of the people of the town are males, 20% are left-handed, which are males, so I times'd them together to get, like, 12%, and then I would take that off of the 21.6% to get whatever was left, which had to be females who are left-handed...' (R31M)

What the student has said so far is in fact correct: having multiplied the proportion of males (60%) by the proportion of those males who are left-handed (20%) and correctly calculated 12%, all he needed to do is to subtract this 12% from the left-handed proportion of the total population (21.6%) to get the proportion of non-male left-handers, which would be 9.6%. But, in the context of the focus group discussion this student realised that this was not the end of the question, even without re-seeing his tree diagram structure. He continued his response:

'... but that gave me the wrong answer to the question, because it's asking for "who are left-handed" but of that, "who are not male." So, instead of females who are left-handed of the whole population, it wanted people who are left-handed out of the females. So I got the wrong answer to that' (R31M).

Student R31M's reflections represent the only fully correct response from Group R. They match up well with comments in the examiners' report: 'many candidates did reach 9.6% but they did not realise that this is the probability of a subject being a 'left-handed not male' and they often gave this as their final response. The key was to link 9.6% with the 40%' (OCR, 2018d, p. 18).

Figure 38 - Ratio Method from Student R31M

Work out the percentage of the people who are not male who are left-handed.

$$M : nm \quad LHM : RHM$$
$$6 : 10 \quad 2 : 10$$

Left handed

$$20\% \times 60\% = 12\%$$

12% are male and left handed

$$(21.6 - 12)\% \text{ are not male and left handed}$$
$$9.6\%$$

↳

9.6%

Source: author's own

Looking at his answer paper (Figure 38 above), it can be seen that student R31M, having underlined the key parts of the question, actually started to use a ratio method, and he was correct in his response, as far as he went. On talking through the problem in the focus group, however, he realised that he had not gone far enough, and that his answer was incomplete.

These are all examples of students with high levels of prior attainment applying different structured approaches to address the demands of the question. All these students gave responses, both verbally and in writing, that appear 'consciously competent' according to the Noel Burch model.

Although questions 2 and 3 on the questionnaire were both concerned with probability, students tended to approach the two questions with quite different strategies. For question 2, 72 out of 95 students (75.8%) used a tree diagram. 38 out of the 72 students who used a tree diagram (or 52.8%) gained full marks, showing how successful this structured strategy was; indeed, only 2 students who did not deploy a tree diagram gained full marks. It is possible to infer that the word 'probability' – which occurs three times in the question – prompted students to recall the tree diagram method. On the other hand, 19 out of 72 students who

used a tree diagram (or 26.4%) gained zero marks: these students knew *to use* a tree diagram, but not *how to apply it* correctly.

The word ‘probability’ did not occur in question 3 on the study questionnaire, although it is a conditional probability problem. A tree diagram method was again one suitable structure for solving the problem, and this method was adopted by 19 out of 95 students (20%). Contrasting this with the more widespread adoption of the tree diagram method in question 2, however, it is possible to infer that, without the trigger word ‘probability’, most students did not think to use a tree diagram. A table summarising the response structures of students is shown below, Table 23.

Table 23 - Group R: Proportion of Response Structures for Question 3

Method/Structure	Partially correct responses (3 marks)	Incorrect responses (1 or 0 marks)	Total proportion of responses
Two-way table	2.1%	0.0%	2.1%
Ratio	9.5%	20.0%	29.5%
Tree	4.2%	17.9%	22.1%
‘Out of 100/1000’	3.2%	4.2%	7.4%
Unstructured	22.1%	16.8%	38.9%
Total	41.1%	58.9%	100.0%

Notes: Group size: 95. There were no fully correct responses for Question 3 in Group R

Source: Author’s own

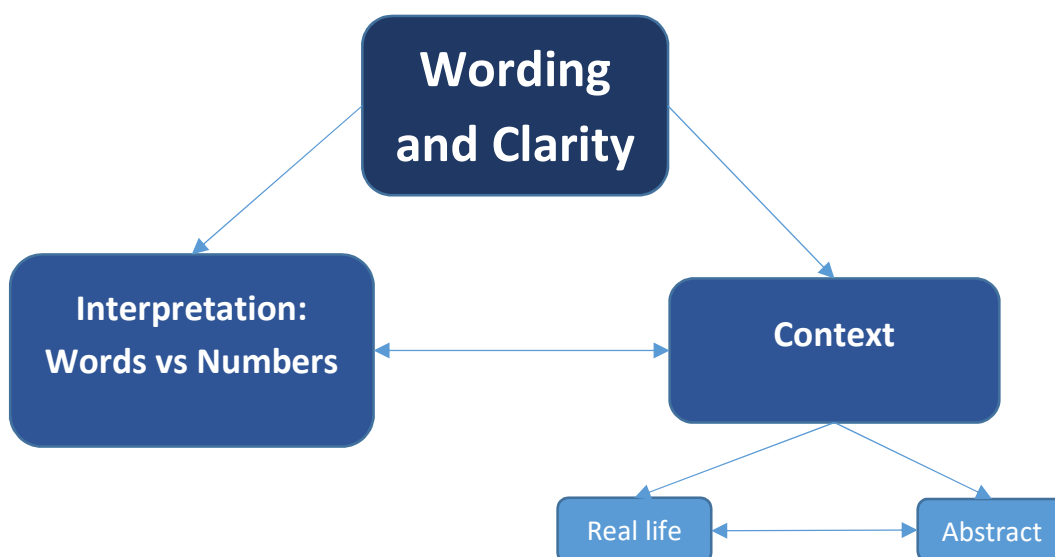
Paradoxically, few students who used a tree diagram gained even 3 marks out of 5, making it proportionally one of the least successful methods. The ratio method was used more successfully: 9 students out of the 15 who used this method gained 3 marks. Few students attempted a two-way table or a method modelling the population as 'Out of 100' (or 1000), although those who did tended to gain some marks. A further 21 students (22.1% of the total) used no structured method – they simply wrote down some relevant calculations – but still gave a partially correct answer and gained some credit. Whereas for a straightforward probability problem (question 2), students in this school appeared to have been taught to use a tree diagram, for the conditional probability scenario in question 3, it was not evident that students had been taught any particular structured response strategy. This question elicited strong reactions and discussion in the two focus groups, because it had disconcerted several students, who did not immediately know how to approach it. Examination papers for mathematics GCSE usually contain one or two questions like this where, presumably, examiners wish to test students' problem-solving abilities rather than recall of a well-learned and practised method.

In summary, students in the main study were able to articulate approaches to answering examination questions that ranged from unstructured trial-and-error attempts to approaches that were much more considered and highly structured. Particular wording in the question – such as 'probability' – often appeared to trigger the recall and application of particular structures and methods. The absence of such wording or direction in a question tended to be met by the application of a whole range of different possible answering strategies, which were more or less structured and, indeed, more or less successful. In discussion, a proportion of students were able to articulate and explain their methods, showing themselves to be 'consciously competent'. A smaller number of students were also able to demonstrate metacognitive awareness, discussing and evaluating the effectiveness of different methods in the context of particular questions.

6.4 “Wording and Clarity” Theme

Almost two thirds of students who were surveyed (62 out of 95 students, or 65.2%) expressed views about ways in which the wording of questions contributed to their perception of the difficulty of questions. Discussions in focus groups also centred on the wording of examination questions. Many students gave responses suggesting that they thought that the wording of a question should present a perfectly clear route to enable them to solve the mathematical problem. A few of these students, indeed, seemed almost indignant when a question did not do this. For others, the way a question is worded offers a discussion point around the extent to which a question is contextualised – rooted in “real life” – or more abstract in its formulation. This is not a binary distinction: there will be a continuum for the contextualisation of questions. Figure 39 (below) shows the interconnected relationship of subthemes within the Wording and Clarity theme.

Figure 39 - “Wording and Clarity” Theme



Source: author’s own

Writing generally, several students felt that the presence of more words tended to indicate a more difficult question, and vice versa: *'the wordier a question is the harder they often are'* (R34M); *'I think when questions have more writing and are longer can also confuse the student making it more difficult'* (R50M); *'I feel like when there's less words I find it easier'* (R87M). In one focus group, student R97M preferred questions with *'less words: cos it's, like, quicker to read and you're straight on with the question.'* Student R87M agreed: *'they're easier to read and easier to get the answer.'* One student summarised this point in his questionnaire response:

'I think that examination questions that are more wordy and less straight to the point are more difficult as you have to find the values to use in the questions within the words describing the reason of the question' (R38M).

This student's explanation makes an obvious but important point: one of the ways in which examiners add to the cognitive load of a question is through the way they give instructions. In many questions that are more demanding, the first step for the student is to work out what the question actually requires them to do; a second step may be to locate and select the values that need to be manipulated; further steps may be to identify and correctly use an appropriate method to solve the problem set out in the question.

6.4.1 Interpretation: words vs numbers

In section 6.3, 'numbers' were considered as one of the question facets to which students directed their attention first; here the 'numbers' are being discussed in contrast to 'words'. For many students, their responses indicated that they thought the cognitive load of interpreting the verbal instructions in the question added materially to its difficulty. This was in contrast to more accessible questions that used numbers in preference to words:

'If a student is given a sum they may know what to do straight away whereas if they are given words they have to infer what to do from the text first adding an extra layer of unnecessary difficulty' (R17F).

'I think what makes them harder is that they are mostly open to interpretation and that you have to recognise the question, know the appropriate method and have to be able to work it out' (R67F).

Student R17F's response indicated that she found the more abstract and compact "language" of mathematics a more effective and direct means of communication than an explanation involving words. Student R67F's use of the phrase *'open to interpretation'* might suggest that she thought there might be a range of possible correct answers, but in the context of her subsequent remarks it appears that she is referring to a choice of method.

For some students, the appearance and presentation of a question could engender an instant loss of confidence: *'when the question looks difficult + it makes you forget the simple rules + maths needed to answer' (R11F); 'worded equations – often don't make sense (issue: the way it's worded, not the math)' (R12F).* Student R29F summed up the differences, as she saw them, between more and less demanding questions. Of a less demanding question, she wrote that, *'it was very straightforward and there was no difficult wording to understand. It was just the question given'* whereas, for more demanding questions, *'I find it difficult to read and understand the question and what the aim is for.'* She did not, however, explain the ways in which the wording made the question, in her view, more difficult to *'read and understand.'*

Many students, like student R17F above, appeared to prefer the transparency of a question that simply stated a mathematical problem. Student R53F stated this most clearly:

'There is often lots of text which can be confusing and takes away from the overall maths of the question. Often Maths questions are more difficult to understand the wording than the actual maths that has to be done for the question' (R53F).

In one of the focus groups, this idea was developed at greater length by four students, referring specifically to the differences between questions 1 and 3. Student R31M began by explaining his negative reaction to question 3 in the questionnaire⁷⁷:

⁷⁷ Question 3: 60% of the people in a town are males. 20% of the males are left-handed. 21.6% of all the people are left-handed. Work out the percentage of the people who are not male who are left-handed. (5 marks)

'The main reason I didn't like this question... is the wording of it. If there's a question where it's mainly numbers, it's quite easy to get it correct, but if it's like, "work out the percentage of the people who are not male who are left-handed"... just the wording of it can throw off all your maths. In one of the tests we had, it was like "what is an assumption that could be made about it?" and it's like you could make many different assumptions... which one could be right, or could be wrong. So I think when it comes down to not being maths any more, that's when it gets tricky.'

This student made a strong claim, that a requirement to interpret the question somehow meant it *'not being maths any more.'* Other students in the focus group picked this up and discussed it further.

'When the question's worded badly, it gets to the point where it's more about you being able to read instead of being able to do the maths; you have to be able to interpret what it's saying, before actually being able to do what it's asking... Like, for the first one, it's really clear: it says, "How much did Reuben pay to hire the car?" You know what you need to do – you just need to do the maths. But if the words are a bit 'off', you have to interpret what it's asking before you can even start with the maths. If that's not your strong point, you've kind of failed already' (R34M).

Another student in the same focus group deconstructed question 3 in the questionnaire, demonstrating that, to him, the maths itself was not the stumbling block:

'I think a lot of people could do this question 3 if it was put out easily, cos the maths of it isn't very hard. But the, like, the last line "who are not male who are left-handed" it can just really throw you off and, like, what are you trying to work out from the question? And I think that's why more people would get it wrong, and it's not just the maths, which is probably what the exam board is probably looking for... If they're wanting to see how good you are at maths (cos it's a maths paper)... if you cannot read it properly or if you misinterpret the question, then I don't think that's what should be assessed, rather than it just being the pure maths skills' (R36M).

The last line of question 3 did indeed seem to *'throw off'* students in this group: not one student found the fully correct solution. In his last two lines, student R36M voiced a view of many students, a feeling that a mathematics GCSE examination should be assessing purely the mathematical ability of students, not their problem solving skills, and that interpretation skills should therefore not be required. Student R49M encapsulated this:

'I think personally with me... words just make it a little – a lot – harder, and problem solving sort of borders maths and completely different like skills; whereas I think realistically you shouldn't really need to like, have specific reading skills, where you can

understand what they're saying, to be able to do the maths question – you should just be able to just answer it without having to sort of like read lower into it cos... you shouldn't have to interpret anything in maths: because it's just sort of like "this is this", it shouldn't be, like, "this could be this or this"... It should tell you what it needs' (R49M).

In this study, the view that *'you shouldn't have to interpret anything in maths'* appeared to be widely held by students, and particularly by male students. In many problems in mathematics, however, a large part of the intended demand is the requirement for the student to work out what the problem actually is, and to bring to bear a suitable method to solve it: it is not simply to process the arithmetic (inherent or explicitly) stated in the question. Many students in this study appeared not to comprehend or appreciate this; or, at least, they would prefer it not to be the case.

Some students rationalised their struggle by reflecting on their own relative strengths across different curriculum subjects; in particular, they contrasted their skills in English and mathematics, sometimes in written questionnaire responses with spelling, punctuation and grammar that also unwittingly illustrated their points:

'It depends on what people personally struggle with. I struggle with questions that involve letters as my brain uses letters in english and numbers in maths' (R81F).

'Personally id say im about average when it comes to maths however im failing english and it has a massive effect with maths since alot of them are comprehension and trying to pick apart what the question is actually asking you, since english if all based off of inference i struggle with it' (R63M, quoted verbatim).

For some students, there was a reaction against wording that they felt was unusual or complex. These students did not, however, pinpoint the source of the complexity in the wording of the questions:

'Factors that make the questions more difficult is when the questions are worded unusual and more complexed as this can confuse the student when answering the question' (R50M).

'I think questions with siffisticated [sic.] words or long questions can make maths more difficult as it can become come confusing and easier to misunderstand the question if it hasn't been read correctly' (R06M).

'The ones where the wording is complex or there are very complex/confusing steps to them that require a lot of attention to the details and preciseness' (R10F).

Student R03F wrote that questions in mathematics are made more difficult by *'complicated wording of questions, added unnecessary information that makes it appear harder than it is.'*

Another student found that the mathematical part was quite straightforward, once he had deciphered the demands of the question:

'To begin with the way the question is worded makes the answer seem much more difficult than it actually was and threw me off. After completion the question seems fairly straight forward' (R51M).

No doubt, this could be said of many questions or puzzles: in hindsight, once solved, they appear straightforward. Nonetheless, it is the propensity of wording in the question to increase demand or cognitive load that the student is highlighting here in particular.

The selection of relevant information may be considered to be part of the problem-solving aspect of a question, but it can confuse students. Student R04F, for instance, considered that *'the wording and adding in non relevant statistics'* in question 3 confused her, *'as I didn't know which statistics to use and which I shouldn't.'* In fact, there were no irrelevant statistics in this question.

For other students, the search for a method could also be complicated by the wording of the question:

'The question was also quite wordy and it can be difficult to know what it actually means. Also usually you can guess sort of what you have to do and where to start but I usually get confused on questions like this' (R35F).

'Understanding the words because if you don't understand what the question is asking you will find it difficult to answer it as you won't know what to do or you could do the wrong method' (R43F).

'The way questions are worded tend to make things difficult for me and can lead me to have to read the question a few times before I grasp the method I am doing' (R44F).

'Information needed to complete the questions are hidden in words and confusing situations' (R56M).

Student R56M's use of the word *'hidden'* is interesting, with its suggestion almost of subterfuge on the part of the examiner. Student R28M went further: *'the question is worded to make us feel like we have been given very little information'*; and he became explicit: *'they are designed to trip us up.'* What motive the student thought an examiner might have in presenting a question in a deliberately misleading way is not clear. A misleading question, with consequently unpredictable levels of difficulty, might be a poor discriminator of different levels of skill among students. Nonetheless, it is important to note the students' perspective, and his imputation of unfairness is one that will be picked up in the discussion that follows this chapter.

A small number of students presented a contrary view, however, suggesting that the wording of a question could actually help them find a method, and this will be explored further in the next sub-section.

6.4.2 Contexts

There is a debate among teachers of mathematics around the desirability of "real life" contexts of questions, and this will be picked up in Chapter 7 (Discussion). In some published studies, gender distinctions are evident, where female students express more of a preference for contextualised questions than males. This was less clear among students in the present study, however. In the focus groups of this study, there were only two female students, and they expressed differing views, so no conclusions can be drawn here about gender preferences.

One female student in a focus group found the real-life context of a question engaging:

'When I look at these ones, I think that they're like real-life scenarios, so it helps to, like, make them understand more... [for example,] when they've got, like, money in them, I find them a lot easier, because they're more like real-life scenarios' (R73F).

For this student, the contextual wording provided additional motivation:

'Sometimes I quite like the ones with words because they're more interesting, so it makes me more likely to do them' (R73F).

Many male students in this study, on the other hand, expressed a preference for questions presented without context. One student was blunt about his preference, although he appeared to understand that his objection might not rest on strong foundations; another student was more analytical in his explanation:

'I'm just going to be honest here. I do not like real-life problems and stuff like that: it just doesn't sit well with my brain at all. And I quite like smaller size questions, because you don't run the risk of not "getting" what the question's asking you to do, which is pretty much my biggest downfall in maths, because I don't actually read the full question' (R77M).

'I think the amount of detail that gets put into the questions can be perplexing for a large number of people, especially if the detail isn't necessary' (R24M).

A female student shared this preference for questions in their abstract form, again seeking ease of comprehension:

'If a student is given a sum they may know what to do straight away whereas if they are given words they have to infer that to do from the text first adding an extra layer of unnecessary difficulty' (R17F).

The key word here is *'unnecessary'*: contextual information may introduce a redundancy effect (Sweller *et al.*, 2011), where additional but redundant material requires processing in the working memory, increasing the cognitive load but interfering with rather than facilitating learning. Other students also recognised this, although they did not clearly articulate how the additional wording made the question more difficult for them:

[Questions are more difficult] 'when there is lots of text and multiple values in the questions that are irrelevant and when the outcome they want for the question isn't worded well' (R53F).

'Factors that make questions more difficult is when the questions are worded unusual and more complexed as this can confuse the student when answering the question. I think when questions have more writing and are longer can also confuse the student making it more difficult' (R50M).

It could be that these students are also referring to the perfectly legitimate ways in which demand in a question is increased by requiring the student to construct the mathematical

problem for themselves from the information given, but their responses suggest that, for them, the wording of the questions confused them, adding additional difficulty that might not have been intended.

Contextual information may be an attempt by examiners to link the abstract mathematical world with the real world as inhabited by students (see Chapter 7: Discussion, below). This may also be an attempt to motivate or engage students. Within the confines of an examination question, however, their attempts may appear somewhat contrived. Student R28M described this dilemma in direct terms, giving three examples of ways in which he felt questions were often made more difficult. Clearly, for him, the added context tended to be irritating rather than motivating:

- *'How they are designed – to trip us up*
- *Not always explicit which methods you must use*
- *Sugar coated through infantilisation, e.g. Timmy has a bag of counters'* (R28M).

By contrast, other questions are presented in a more abstract form, and many students appeared to prefer this, particularly when the question was also shorter.

[More difficult questions:] *'I think when lots of words are added like in Q3 as you've got to figure out what to do from the words. Where[as] in all of Q6 you're given a sum and a one word instruction so it's easier to understand'* (R17F).

In the second focus group, a debate opened up between the two female students. They heard some of the male students state categorically that they preferred *'smaller questions – there's less, like, to read'* (R97M) – by which this student meant *'less words... quicker to read and you're straight on with the question.'* This was agreed with by another student, *'they're easier to read and easier to get the answer'* (R87M). One of the female students then offered a different perspective:

'When there's, like, a long-winded algebra question with a lot of marks and no context, I think they're quite hard' (R73F).

[Researcher:] Go on, tell me why.

'There's just a lot more to do and it's like, down to you, because there's not much help with the questions' (R73F).

[Researcher:] Ok. So, it's "down to you": the question doesn't help you much... at that point, do you rely on what you've learnt in a lesson, a method maybe you've learned?

'When you know what you're doing with these ones it's fine, but when you don't it's not as easy, but with these ones you kind of get help from the words in the questions' (R73F).

This student recognised that the wording of questions could actually help – rather than hinder – the student, particularly if the student did not instantly realise what they needed to do to address the question. Her fellow student in the focus group, however, suggested that the wording could obscure the task:

'I think I do prefer the ones with just numbers, just cos sometimes the words, I can get quite confused with, like, which one's asking us what sometimes. I don't mind the words, but if it was just like the shorter number questions, I think I prefer to do them, cos I feel like I understand them more and... if I quickly do them then I'll be able to spend more time on the 'wordy' questions' (R96F).

[Researcher:] Now that implies that you think that either the 'wordy' questions are harder, or that you'll find them harder... which is it, d'you think?

'I think – with the 'wordy' questions, I think it just depends on the questions... If it's like normal little 1-markers then I think I'm ok, but normally... where the 'wordy' questions start, like, being 4-markers, 6-markers, then it can get, like, you need to spend more time on them' (R96F).

For this student, then, questions with more words tended to be associated with higher numbers of marks. She expected these to be more difficult, but it is not clear that it was necessarily the wording in itself that provided the additional demands.

Although one student in the first focus group preferred a shorter question style – *'I would say [it] would be more understandable to me, because I don't need to worry about reading any words'* (R31M) – other students described how they found additional wording helpful at times, compared with questions that were more abstract or pared-down in their approach:

'I think lots of people might struggle with [question] 5 because there's less to go off, and... if you don't know what's going on you won't have a clue because all it says is "solve" for number 5 (a)... but on 4... you've got lots of words to help you understand,

you've got context given at the start of [question] 4 before even any questions start, so it might be easier to get an answer if you're not quite sure about, if you're like not quite sure of what to do' (R36M).

For this student, a shorter question offered 'less to go off,' compared with the scaffolding effect of 'lots of words to help you understand.' Another student in the same group was even more explicit about the potential assistance and assurance given by additional wording:

'I feel like, in questions where there are lots of marks, the more information given, word-wise, even if it makes it more open to interpretation, does make it more accessible to lots of people, cos they'll be like, right, well, towards the start it's going to be the first couple of marks, towards the end's going to the last couple of marks and they do manage to, like work through it' (R31M).

This student used the words of a question to help him chart his way through its requirements. In this way, the student regarded the additional information and wording of the question as decreasing the level of its demands, making it 'more accessible to lots of people.'

A student in the second focus group referenced the scenarios that were given to him by his teacher. He saw the benefit of these (and the link to repeated practice):

'Our teacher, like, gives us different forms of questions, which gives different... scenarios and, er, the more them that you can learn how to do like the easier it becomes... by different types of questions on the same topic' (R77M).

Contextualised questions, in this questionnaire, were not longer than more abstract questions, and nor did they carry more marks. Nonetheless, many students referred to finding contextualised questions more complex and challenging than more abstract questions. The use of a context, more or less related to a real-life scenario, appeared to motivate some students but confuse or irritate others.

6.5 "Memory, Practice and Familiarity" theme

For very many students in this study, memory, practice and familiarity were closely associated with question difficulty. Half of the students surveyed made comments that linked some aspects of memory, practice or familiarity with question topics to their perceptions of difficulty (47 out of 95 students, or 49.5%). Twice as many female students as male students made comments on this theme (33 females, 16 males; some students made more than one comment). There were 25 mentions of words associated with memory or remembering, and a further 16 mentions of words to do with forgetting. 10 students talked explicitly about practice. Students in the two focus groups also referred to practice and familiarity. As might be expected, the majority of student comments made a simple link between their familiarity with a topic and the facility they were able to deploy when tackling a question. Broadly, then,

easier = more familiar, well remembered, practised frequently

and, conversely,

more difficult = unfamiliar, not well remembered, not practised consistently.

Students' views were richer and more nuanced than this simple distinction, however, and they are explored in more depth below.

6.5.1 Memory

Beginning with the semantic (surface) level of the "memory" sub-theme, one student summarised his view of demand and difficulty, relating these concepts directly to degrees of memory and familiarity:

'I don't think that there is anything that makes a maths question difficult, because if I know and remember the correct method and formulae to figure out a question, then it will not be difficult, however if I have forgotten what steps to follow to get the correct answer, then I will not know what to do' (R58M).

This response, illustrative of a number of student comments, suggests that the student does not feel in control in meeting the demands of a question – it is almost fatalistic in its

suggestion: he may know and remember the correct method; or he may not. It is simplistic in its implication that it is the level of the student's familiarity with the topic and its methods alone that determines the difficulty of the question. At face value, the student is unconscious of the examiner's role in creating difficulty: his response does not analyse any features inherent in the question design, but documents merely his own role in remembering what he has been taught. If this were true universally, examination questions would be of limited use as discriminators between students, and poor predictors of their future performance, because they would test only the students' levels of memory and recognition. However, the comparison of questions on the same topic in the questionnaire (if nothing else in his experience of exam questions) might have shown this student that it is possible to present a range of questions on the same topic, in all of which he might be very familiar with the necessary methods and formulae, but the questions could pose different levels of demand and complexity.

It is evident that some students do think in this simplistic way: for them, examinations are, at least in part, memory tests. Students wrote, for example, that,

'Sometimes people just don't remember stuff as well as other people, then its just a memory test, not a knowledge test' (R26F).

'I think you're at a disadvantage if you can't remember the formula you need to use, that can lose you a lot of marks right of the bat' (R62F).

[It's harder] *'when you forget how to work the questions out' (R76F).*

The first two responses infer that a memory test would be arbitrary and unfair, whereas the result of a knowledge test would have more validity, in the students' eyes. The third comment hints at the link between repeated practice and memory: thinking about the working memory model, *'you forget'* implies that you were taught how but didn't practice retrieval sufficiently to drive and encode the method into the long-term memory so that it could be securely retrieved in an examination.

Question 6, in particular, which required students to remember and correctly apply the formula for finding the circumference of a circle (this was not given in the question paper), drew a number of comments, typically, *'I forgot how to work out the circumference for a bit so I was stuck'* (R66M). It is true that questions requiring simple recall test whether a student has securely learned a formula, but do not discriminate between different levels of skill: this is, presumably, not their purpose.

Only a small number of other students wrote in this vein, however. Most other students' comments show they think the interaction between memory and difficulty is more complex than this; for example:

'I believe that most of the questions are skill based, however those that require memorising formulae I think is based on your ability to remember them' (R44F).

This same student recognised that memory is not a stable or consistently reliable function, commenting for one question that,

'I forgot how to do the question for a few seconds, but then figured it out and it wasn't difficult' (R44F).

For this student, the question was impossible at first, when she had forgotten a suitable approach. Tantalisingly, she did not explain whether she subsequently remembered the method she had been taught, or whether she worked out a suitable method, perhaps by trial and error; possibly it was a combination of the two.

The specific effect of exam pressure on their facility to recall things they had learned was referenced by two students. Student R29F wrote that *'I personally find it alot of pressure in exams and forget alot'*; student R41M agreed: *'[in exams I] just forget. Crumble under pressure.'* In the review of literature (section 3.1.1 above) the action of stress on reducing working memory capacity was noted. The role that memory plays in increasing or decreasing the experienced difficulty of examination questions for students, then, is not a stable one. This will be discussed more in the following chapter.

6.5.2 Practice and familiarity

Repeated practice was a sub-theme mentioned explicitly by 12 students. One student, for example, commented that questions 5 and 7 on the questionnaire were easy,

'Because we've learned it at the start of secondary [school] and have constantly had practice by either revising it in lesson or using it to solve other questions' (R17F).

Other students made similar points, stressing recent practice and consequent familiarity. One said that she could manage a question because she *'had done it in class not too long ago'* (R75F); another wrote that she was *'familiar with the type of question and knew how to work it out with the correct method'* (R86F). This student also cited how repeated practice influenced her judgement about difficulty between the methods expected by different questions on the same topic:

[Question 7 was easier] *'because I had more practice in question 7 and was confident in what I needed to do'* (R86F).

Other students made the link between repeated practice and their consequent confidence.

Confidence is connoted with their use of words such as 'just', 'simple' and 'easy':

'In class we have been doing this topic therefore I have had practice and didn't find it too difficult' (R74F).

'We do this alot in lessons as it just something simple I can pick up as well' (R71F).

'Probability is easy as we have spent a lot of lessons on it' (R87M).

'We have been learning how to [approach] probability and factorising for a long time. It is just easy' (R19F).

Conversely, students reported that a lack of familiarity increased the difficulty they experienced. One stated that *'I haven't done it or recapped it in a long time so I forgot'* (R70M); another said, *'I am not familiar with this type of question therefore was unsure how to answer it'* (R74F). It is a fundamental feature of the working memory model that repeated practice

encodes schemas into the long-term memory. Although they had not been taught this model, these students clearly understood one of its tenets.

Students' responses indicated that it was not just topic familiarity that helped them, but also actual practice with the methods they needed to use:

'I have had a lot of practice on those types of questions so I remember how to do it really easily, and I recognised the method that I had to use straight away' (R67F).

'It is the easiest method to do and to remember. We have been learning about it recently... Simple method, practiced a lot' (R57M).

Increasing familiarity, through repeated practice, was mentioned by one student as her key to success:

'I think it if just keep practising it, then it'll just sink in. And it tends to be working, so sometimes I think it's ok' (R96F).

Another student gave an example of how repeated practice, across questions of increasing demand, helped him to improve his learning:

'I think the main way that you can learn and get better is to do questions and see where you go wrong. Like, I used to get negatives wrong all the time, so by seeing that, I'm able to check, have I got everything correct when it comes to that. I think mainly just practice and doing questions that you wouldn't usually be comfortable with' (R31M).

A few students were able to go further in their evaluation of their own understanding, suggesting levels of metacognition. One explained that a question could be difficult,

'When knowledge of content needs to be put into a context that has not been explicitly practised in lessons' (R36M).

Asked further about this response in a focus group, the same student added,

'I think if you're learned... if you've done a topic and you've done the content for that topic, if you go to the higher end of the spectrum for that, like the difficulty of questioning for that topic, I don't think it could be too hard, because you've been taught how to do the simple, so you've just got to apply the... this is going to be a harder question for that. So you use your skills from doing the easy part, and then you, you try your best to just get anywhere near to the answer' (R36M).

This student recognised that part of the demand in an examination question is the expectation for a student to apply their prior understanding of a method and topic to a new context. This student also demonstrated his resilience: he expected to struggle to *'get anywhere near'* to a correct response.

Schmidt and Bjork (1992) investigated the role of repeated practice and training, for example across a variety of questions and applied scenarios, to enable a student to tackle unfamiliar questions in a known topic. This thread was explored in the focus groups in this study.

Sometimes repeated practice could appear dull to a student, but the value of that practice was still recognised. For example, one student in a focus group referenced the value of repeated practice, to the point where students instantly recognise common formulations of questions:

'We've practised these types of questions lots in lessons – but we've done them, like, to death, so we know how to do these ones' (R31M).

This student's description of repeated practice calls to mind the model of working memory and long term memory (summarised in section 3.1.1), where maintenance rehearsal in the working memory leads to encoding into the long term memory, so that understanding that has been learned may be retrieved when needed in future. In the other focus group, a student explained that her teacher would mix things up a bit in lessons, to create variety. Talking about her maths lessons, she said that,

'I think it's good how they, like, incorporate different types of questions so it's not all, like, just small questions – you've got the bigger questions and they're not, like, maybe all the different types of questions. If it's a certain topic, they'll change it around a bit, the way the question's formed or something like that... it just makes it different, so you're not doing the same repetitive thing each lesson' (R96F).

It is not possible to know from this student's statement whether she thought the teacher would have been varying the questions for variety's sake within the lesson, or because she realised that the teacher understood that examination questions are also presented in different formats, and the student needs to be prepared to encounter a variety of different presentations without being put off by the differences. This latter understanding of teaching

intentions was a little clearer to another student, R31M. He explained, in relation to question 4 on the student questionnaire⁷⁸, that he knew at once what to do

'With the words, "Write down the value of the car when it's new", because we've gone over the fact that in the equation V equals whatever times whatever to the power of n , the first [number] is what it would be new... we've gone over [that sort of question] in class lots, so we could apply that to any variation of that question because we've learned the skill for it' (R31M).

[Researcher:] Tell me more about that.

'So for instance, this question is about the value of the car when it's new but if another question came along that was asking what's the value after 3 years, so because we've done the ins and outs of this question, we've learned how to answer it, we'd still be able to do it, even though it's like a completely different question' (R31M).

This student demonstrated a secure grasp of several essential mathematical skills, and showed that he also comprehended the intentions attached to the verbs used to trigger the deployment of these skills in addressing examination questions. He explained that

'If it's asking you to "solve" something, we know we need to find the value for it. If it's asking us to "prove" something, we know we'll have to use algebra to show that without a shadow of a doubt something is something. So instead of "showing" it, like, using an example of, if you say that an even number plus an odd number will always be odd, it's different showing it like $3 + 2$, as it would be to say, like, $n + (n+1)$. It's just different ways of answering them: once you learn those, no matter what the question is, you'll have a good shot at it' (R31M).

Like student R31M, a few other students understood that the repeated focus on particular questions, methods and topics in their mathematics lessons was about more than preparation for examinations; it was about drilling them in skills that are foundational for other topics (and, indeed, for other subjects):

'We cover this topic a lot in maths. Also this is a required skill in other topics so you need to know how to do it well to get higher stakes questions right. So there is more practice on this' (R35F).

A discussion opened up between students in the first focus group around the value of repetition and variety in maths lessons. This was an interesting debate that touched on aspects

⁷⁸ Question 4: The value of a car, £ V , is given by $V = 16,500 \times 0.82^n$, where n is the number of years after it is bought from new. a) Write down the value of the car when new (1 mark)

of challenge and student motivation as well as ways to encode learning in the long term memory and therefore reduce the difficulty of examination questions:

'I would say that in our school we've got a quite good maths department, a really good one actually, and they're able to engage everyone quite well. I think the ways that they do this is by going through with the class together and then helping students individually that are struggling. But also to give those that are further on challenges, so that they can keep interested, instead of doing the same work... cos the way that I do something would be different to the way that [student R34M] does something' (R31M).

'It's like about variety in what you're doing, so that, say there's a kid who doesn't understand maths and doesn't enjoy it, and they just do the same lesson, basically, for two years straight, they're just – they're going to lose interest, then once you've lost interest, you're not going to enjoy it again. So, if they [the teachers] keep changing the lessons, they might find something that engages, like, a struggling student, and that might help them later on, like, cos they want to learn more' (R34M).

A third student in this focus group further extended the discussion, into a consideration of elements of challenge and reward; his comments will be reported in the final section of this chapter. These students, who were in the top stream for GCSE Mathematics, were able to articulate a mature understanding of the value of repeated practice in their lessons. In their views, repeated practice helped them to remember methods and apply them in different contexts, and it kept them engaged and motivated in their learning. This theme of motivation, as discussed by students, will be analysed in the next section.

6.6 “Motivation” theme

The fourth theme, “motivation”, was developed at a deeper level from students' responses in the two focus groups; students had not been asked about their motivation in the questionnaire. Some students reported positive motivation, in terms of enjoying tackling questions in mathematics, and there was also an element of negative motivation, in the form of a fear of failure. Jackson (2010) has suggested that fear can be a useful lens through which to view many aspects of education, and that students tend to fear academic and/or social

failure. Students' approaches to examinations and examination questions lend themselves well to this analysis.

Students in the focus groups admitted that they could find examination questions '*intimidating*' (students R34M and R31M), although a lower mark allocation could mitigate this:

'In [question] 4 you can look at it and go, like, "ah, there's loads of words," but if you look and say, "ah, there's just 1 mark," it just like takes all the intimidation out of the question' (R34M)...

[Researcher:] I like the way you said it "takes the intimidation out of the question" ...

'...Yeah. Cos, a lot of people, like, they might look at question 4 and they'll see, oh it's a maths question but they'll see like 6 lines of words, but seeing it's 1 mark it'll make people think it cannot be hard, it's not going to be difficult, so it'll calm them down, maybe make them just read through' (R34M).

The sense of panic in the student's second response is almost palpable. His words suggest he is speaking of an inner voice that he might use in an examination, consciously helping himself to keep his fear and anxiety under control.

At an opposite emotional pole from intimidation is fulfilment. For several students in this study, mathematics brought real pleasure and a sense of satisfaction. These were students who were motivated by their own success, and by their ability to help others. One gave an explanation of how his motivation developed, from his time in primary school, at first based on games designed to help pupils remember their times tables,

'And ever since then I've always loved doing maths, like, I just love anything that involves like working out numbers and anything like that, because it has a definite answer, whereas in things like science or like English it doesn't always have a definite answer within it, so I like it for that reason' (R51M).

This student picked up and developed a strand evident in several other students' questionnaire responses, that they liked the definite feeling of knowing that they had found the correct answer to a question in mathematics, a feeling that was more elusive elsewhere in their school curriculum.

For other students, the satisfaction of learning new skills was motivating, particularly as they could see that they had grasped new concepts and increased their levels of knowledge and skill:

'Learning new things was best for me. Like, especially with like maths, just sort of doing things you've never done before, learning completely different things from what you were used to, sort of helped me... like, if I don't know something and then I know it, I'm obviously getting better at it, which means if I just keep looking at things I don't know then eventually I'll learn them and get better at maths' (R49M).

For this student in particular, but also for others, there was an element of challenge that he enjoyed:

'If you give someone something that's not going to challenge them, they go, oh that's quite boring and they're not going to enjoy it, but if you give someone something that's at the top of their ability, that they can, like, they can learn how to do but they don't already know how to do, that'll challenge them to do something, like, harder than they're actually doing. So if you actually, like, sort of, push people out of their comfort zone when it comes to answering questions, you'll be, you'll be more likely to learn if you're answering harder questions' (R49M).

This student is articulating something very close to Vygotsky's ideas about questions located in the student's Zone of Proximal Development (summarised in Chapter 3). Other students picked up this point and developed it, showing that challenging questions also provided stimulus and interest, as well as the opportunity to synthesise their understanding. This enjoyment of challenge was not restricted to males; here is a female student speaking:

'Maths is one of my favourite subjects, because I've always, like, quite liked it a lot. But, um, when we do, like, challenging questions I quite like those, like the long ones, cos there's, like, a lot of things you can apply it to, so it's like putting all your knowledge together' (R73F).

[Researcher:] Ok, so you quite like it when it's hard, when it's more difficult? [...] Why d'you like that?

'I dunno, I just like always being challenged by it. Like, it's just something different!' (laughs) (R73F).

Some students extended their thinking to include the crucial role of the teacher. They saw the teacher's role as extending beyond being merely instructional, and more to providing challenge and variety.

'I think that's also the job of the teacher to challenge the students and engage them by giving them difficult questions, but instead of telling them the answer when they struggle, to just give them a pointer or give them a hint to one part of it... and you go, ah, now I need to find that out. And it can be, it can be quite motivating if you then manage to go and do the question on your own, just with that little bit of help instead of being given the answer' (R31M).

'I think it would be good for teachers to give a little, like, a challenge question on the board for questions further on in the topic that they think you could maybe achieve, if you really, like, think of it... just to challenge your brain in thinking, what haven't I done yet and what could link to what I've been doing in this last couple of lessons' (R36M).

'I think it's good how they incorporate different types of questions so it's not all, like, just small questions... If it's a certain topic, they'll, like, change it around a bit, the way the question's formed or something like that. It just makes it different, so you're not doing the same repetitive thing each lesson' (R96F).

So part of the role of the teacher, according to these students, is to provide sufficient challenge to maintain the interest and motivation of the students; to provide a suggestion or hint when a student is stuck, but not to supply the answers because that could be demotivating. In providing variety in the type and presentation of the question, whilst not necessarily introducing any new concepts or topics, the teacher is also enabling the student to rehearse and reinforce their understanding, strengthening the construction of schemata, according to the Baddeley and Hitch working memory model (Baddeley and Hitch, 1974; Baddeley, 2002), and avoiding cognitive overload.

Finally, students explained how they found being able to succeed in tackling more challenging questions was rewarding for them, both intrinsically and because of the positive social engagement and return they obtained from being able to explain to others. This was shared by male and female students.

'The more challenging questions are the most interesting and, er, I don't know how to word this, but when they're really, really hard and not many people get it, there's a

select few of people who will, like, get it, like, the first try. And I prefer it when it's like that... I have the ability to, like, help those around us, [student R87M] for instance: [he] quite often asks us for help. I enjoy that – I enjoy actually helping people: that's fun to me' (R77M).

'I think sometimes it does help to, like, explain things to other people... it shows that you understand and it's like telling yourself you understand... If we're going through a question and I sit next to my friend and we're going through a question it, like, helps me understand if we're going through it together... cos it shows we both know something about it... To get another person's perspective's good as well... cos I think differently to how my friend thinks – like, everyone thinks differently, so when you look at it from a different point of view, that sometimes helps you' (R73F).

These last two perspectives reinforce the social side of learning and resonate with perspectives reported by Goodenow and Grady (1993): students are motivated from learning together, and from helping one another, and both their learning and their friendship are strengthened. This is the antithesis of the fear of academic and social failure that Jackson (2010) described: here students are describing how academic success can also engender social success at school.

6.7 Conclusion of the Main Study

In the main study, student responses were particularly deep and interesting, giving insights into their approaches to examination questions, the ways in which they apply pre-learned structures and methods to unfamiliar contexts, and the ways in which their motivation is affected by encountering more or less demanding examination questions. Through a process of reflexive thematic analysis, themes were developed around students' attitudes to the wording and contextualisation of examination questions, and the ways in which memory, practice and familiarity effect their experiences of difficulty. These rich responses resonate with and also go beyond some of the themes developed in the pilot study. In the discussion that follows, students' insights discovered in the pilot and main studies will be brought together in the context of literature reviewed earlier in this thesis.

Chapter 7: Discussion – Telling the adventurous story

In this chapter I arrive back home, reflect on my journey and “tell the story of my adventures,” in the form of a discussion of the pilot and main study findings and their implications.

This thesis set out to examine the ways in which examiners create and manipulate factors relating to demand in GCSE mathematics questions, how these cause difficulty for students, and how students experience and comprehend these factors. Examination questions are the means by which examiners sample the domain of the students’ expertise, knowledge and understanding. The marks and test scores that ensue are ‘incomplete measures, proxies for the more comprehensive measures that we would ideally use but that are generally unavailable to us’ (Koretz, 2008, p. 9). Within schools, and within the commercial enterprise that educational assessment has been for many years, it would be paradoxically easy to overlook the students themselves, and not to listen to or consider their views. In this context, this thesis posed a single research question:

How do students experience and comprehend demand and difficulty in GCSE mathematics examination questions?

Through the discussion in this chapter and the conclusions arising from it, I take the rich database of student evidence that I collected through survey questionnaires and focus group interviews, and analysed using a reflexive thematic analysis method, and I attempt to construct and develop meaning and understanding from it.

7.1 The ‘Big Q’ approach – qualitative methods within a qualitative framework

In Chapter 3: Methods, the qualitative research paradigm was introduced and explained. This qualitative paradigm has been applied in this study through a process of reflexive thematic analysis, following the guidelines of Braun and Clarke (2022). Sometimes called the ‘Big Q’ framework (after Kidder and Fine, 1987), the qualitative paradigm differs from a quantitative

approach in several important ways. The purpose of the research is to focus on understanding situated *meaning* and through this to gain a better sense of the students' lived experience of demand and difficulty in the GCSE examination questions they encounter. It is important to recognise, in this phenomenological approach, that the students' experience is not fully knowable, and that there may be many different partial "truths" to be told and explored through this research. It is a strength of the qualitative paradigm that the researcher is embedded within the world they study, as a situated interpreter of meaning and a subjective storyteller, bringing the experience of the students to life through their own words, and investigating it by relating it to and comparing it with the insights of other educators and researchers. The inevitable subjectivity of the researcher is not just acknowledged: it is valued and regarded as an asset (Braun and Clarke, 2022, p. 6). This subjectivity is continually interrogated, however, honed and refined through a continuous process of the researcher's reflexive engagement with their material and their themes. The purpose of data analysis in this qualitative paradigm is to focus on text and meaning. The richness of smaller, more immersive samples, is valued – such as the discussions between individual students in focus groups – since they lead to opportunities to gain in-depth understanding of the students' actual experiences. The present study, following this 'Big Q' framework, therefore aims not to be a stepping stone towards complete or perfect understanding, but to become part of a rich and multi-layered tapestry of understanding of the relationship between students and the institution of high-stakes assessment in which so much of their schooling takes place.

This qualitative research paradigm is different in many key ways from the quantitative paradigm in which much research takes place (including an increasing amount of research into educational assessment). Aspects that might be regarded as weaknesses in a quantitative paradigm – such as researcher subjectivity, the focus on meaning rather than data, and the lack of an initial hypothesis to be tested – are valued and celebrated within this qualitative methodology and paradigm.

When discussing empirical studies using inductive methods, where no initial hypothesis is advanced and where thematic analysis is used to create models of understanding, it is not possible to give definite predictions in the way that, perhaps, a traditional scientific investigation might hope to do (Bassey, 2001). Similarly, when dealing with subjective psychological concepts, such as students' views of demand and difficulty in examination questions, relative – not absolute – positions are adopted. The concept of 'fuzzy predictions' (Bassey, 2001) has, therefore, been adopted by this study as a way of encapsulating the findings of the empirical study and communicating them to possible future users. Elucidating further upon fuzzy predictions, Bassey explained that,

'A fuzzy prediction replaces the certainty of scientific generalisation ('x in y circumstances results in z') by the uncertainty, or fuzziness, of statements that contain qualifiers ('x in y circumstances may result in z') ... Fuzzy prediction invites replication and this, by leading either to support of the statement or its amendment, contributes to the edifice of educational theory' (2001, p. 5).

In this chapter, Bassey's fuzzy logic is employed, couching conclusions and findings within conditional language ('may' instead of 'will,' for example). The extent to which the findings and conclusions reached in this chapter can be generalised are, in addition, qualified by specifying the circumstances in which they might be most likely to operate. It is hoped that future researchers, building upon the unique contribution to the fabric of existent academic knowledge that this study presents, may wish to investigate further the extent to which the conclusions reached here operate within other educational circumstances.

7.2 Revisiting validity

The concept of validity is much discussed around assessments. This study has followed the consensus definition of validity (Newton, 2012), that validity is a property not of the test but of the interpretation that is brought to the result of the test. Kane stated that 'an interpretation is said to be "valid" if it is supported by appropriate evidence' (2016, p. 198). Almost regardless of the interpretations that will be made of the result of a GCSE mathematics examination, the validity of these interpretations is likely to be stronger if there is a clear and evident link

between the expertise of the student and the results they obtain. For validity to be at its strongest within the examination system, the real system would strongly resemble the simplified world set out in the Introduction – the Utopia Ltd Examination System – there would be no ‘unexpected difficulty,’ so that ‘demand,’ intended by examiners, would translate directly into ‘difficulty,’ experienced by students. Since students would approach the same questions with different levels of expertise (understanding and prior learning), this would allow examination grades to be seen purely as a function of the ‘demand’ intended by examiners and the expertise of the students. Perfectly valid inferences, about the students’ expertise, knowledge and understanding, could then be made from the examination grades. The inferences could then become the basis for other uses of examination results, such as predictions of future performance. There would be no ‘noise’ in the system, only ‘signal.’

In the real world experienced by actual students sitting actual examinations and facing real questions set by real examiners, however, there are several sources of noise, and the simplified model has to be adapted to take account of these sources of noise – or unexpected difficulty – that have been revealed. Because it interrupts the pure functioning of the examination system as a perfectly valid measure of the expertise of the student, unexpected difficulty poses threats to the validity of inferences made from interpreting examination results.

7.3 Students’ experiences of fairness in high-stakes assessments

There is an important aspect of fairness and social justice in this approach to assessment. Observing that ‘high-stakes assessments are a common stamping-ground for debates about fairness,’ Nisbet and Shaw (2020, pp. 2-10) explored six different senses of fairness, and summarised that the two most useful senses for understanding educational assessment might be an “implied contractual sense of fairness” – ‘something is fair if it meets the legitimate expectations of those affected’ – and a “relational sense of fairness” – something is fair if it treats similar cases alike. McArthur (2018, p. 195) stated that standards, guidelines and

procedures should reflect ‘the lived realities of assessment, and students’ future lives’ in order to be fair to all test-takers. A more inclusive, proactive role for students in engaging with assessment practices has been encouraged by a range of authors (see, for example, Entwistle, 1991; Orr, 2010; Carvalho, 2013; Boud and Soler, 2016; Black and Wiliam, 2018), but this call has not been answered so far by the actions of examination boards. If important stakeholder voices go unheard, it is hard to argue positively for the fairness of the assessment process. Claims of unfairness around high-stakes assessments are not hard to find – for instance, the ‘scandal’ (Guardian, TES, 2012)⁷⁹ over GCSE English grades in 2012; and ‘dismay’ over the content and accessibility of SATs Reading tests in 2023 (DfE; Guardian, 2023)⁸⁰. As has already been seen in the findings of the preceding chapters, students often expressed their feelings that aspects of their examination questions struck them as potentially unfair – words such as ‘trick,’ ‘trip’ or ‘trap’ have been quoted – where the wording of a question was experienced as an unpleasant surprise and the question therefore failed to live up to the students’ reasonable expectations. These issues will be explored more in the section on unexpected difficulty below. Nesbit and Shaw concluded that fairness was a ‘necessary but not sufficient condition for validity’ (2020, p. 147) in high-stakes assessments, and that, since fairness is a continuum rather than a binary concept, no assessment can be completely fair, but that it is reasonable to ask how an assessment could be moved further along this fairness continuum. Understanding the views and lived experiences of students is, this thesis suggests, a necessary step to making examination questions fairer.

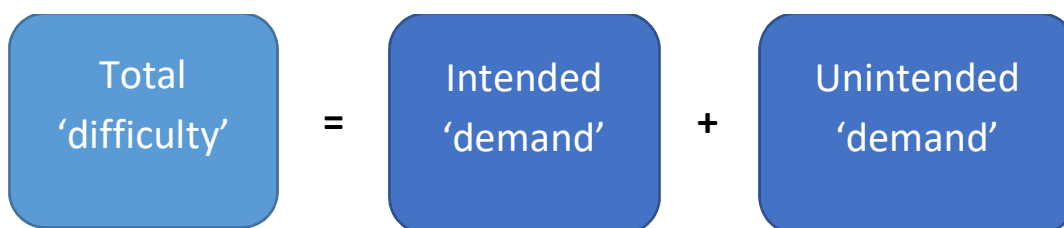
⁷⁹ <https://www.theguardian.com/education/2012/oct/11/45000-resit-gcse-english-exams>;
<https://www.tes.com/magazine/archive/gcse-english-2012-grading-scandal-evidence-schools-were-right-all-along> both accessed 05.07.2023

⁸⁰ <https://educationhub.blog.gov.uk/2023/05/18/sats-english-reading-test-were-the-year-6-tests-more-difficult-in-2023-than-previous-years/>;
<https://www.theguardian.com/education/2023/may/11/headteachers-express-concern-over-sats-amid-claims-a-paper-left-pupils-in-tears> - both accessed 05.07.2023

7.4 Recapping the conceptual model relating demand and difficulty to cognitive load

In Chapter 3, a conceptual model was presented, in which the total question difficulty (total cognitive load) as experienced by the student, was made up of the demand factors intended by the examiners (intrinsic cognitive load), plus any other sources of difficulty (extraneous cognitive load). This model was presented in diagrammatic form (Figure 6) and is here reprised as Figure 40.

Figure 40 - *Conceptual Model of Cognitive Load*



Source: Author's own

Revisiting this conceptual model following the literature review and the reporting of the results of empirical studies, it is now possible to understand much more about what makes up intended demand, and what contributes to the unintended demand.

7.5 Understanding demand and difficulty

In Chapter 2 a set of definitions was discussed with regard to what professional educators – the assessment community of teachers, examiners and researchers – understood by the concepts of demand and difficulty when discussing GCSE examination questions, as well as why these concepts were important.

The definitions and understanding of these key concepts are now summarised. These definitions are based on published research and professional discussion, and it might be beneficial if they formed the basis of a professional understanding that is shared more widely,

so as to avoid what can be unhelpfully loose use of this technical terminology. Bringing together both the definitional understanding of previous authors such as Baird *et al.*, (2009) and Pollitt *et al.*, (2007), and from the information generated by this study, this study advances a more nuanced definitions of both the concepts of demand and difficulty and how they are used.

Demand is a theoretical concept. It is qualitative by nature, meaning that it is best explored in non-numerical ways, searching for understanding through description and explanation (Cohen *et al.*, 2018). Demand is determined by examiners. The following points exemplify demand in the context of examination questions.

- Demand is the cognitive load imposed by a question. It is a combination of the complexity of the topic and its abstractness, and of the levels of cognitive processes required by a question in terms of the knowledge and resources the student is required to supply, and the steps they have to go through to address and answer the question.
- Factors relating to demand ought to be in the control of examiners; they ought to be able to predict, with a reasonable degree of accuracy, the nature and level of demands they are imposing within a question. The principal sources of difficulty, as students experience it in answering examination questions, should be this intended demand.
- Intended demand is fixed for a particular question. It does not vary for different students.
- There may, however, be additional demand factors that examiners have not intended or anticipated, which nonetheless create difficulty for some or all students. This is what is meant by 'unintended demand,' and this may vary between students.

- The total demand imposed by a question is the accumulation (sum) of intended and unintended demand.
- Demand is essentially qualitative – it can be described, but it is hard to measure absolutely or objectively. Nonetheless, the level of demand imposed by different questions can be compared and evaluated.

Whereas demand is determined by examiners, difficulty is experienced by students. Difficulty will vary from student to student. It is quantitative by nature, in that it can be measured and evaluated in numerical terms. The following points exemplify difficulty in this context.

- Difficulty is measured quantitatively as the performance of students answering the question. There is an inverse relation between marks gained and difficulty: if more students answer a question correctly, then the question is by definition less difficult.
- Individual students experience varying levels of difficulty in a question, depending on their levels of preparation and expertise.
- In order for an assessment to be as fair as possible (within the ‘implied contractual’ and ‘relational’ senses of fairness outlined by Nesbit and Shaw, 2020), the main sources of difficulty in an examination question ought to be the intended demand factors planned by the examiners. This would strengthen the validity claims of the examination process.
- There may be a difference between ‘perceived difficulty’ and ‘experienced difficulty’ for students. It was found in the pilot study that students had clearer and more accurate views of the demands of a question once they had attempted to answer it. Some students in the main study articulated this cogently, describing the difference between the higher levels of ‘perceived difficulty’ as they first approached a question, and the lower levels of ‘experienced difficulty’ once they had comprehended what they needed to do.

- Sometimes, students experience aspects of a question in ways that may not have been intended or anticipated. Thus, there may be a gap between examiners' intended demands and students' experienced difficulty. These sources of unintended demand create additional, unpredictable sources of difficulty. This unpredictability may threaten the validity of inferences made from examination marks and grades. The perceived fairness of the assessment may therefore be compromised.
- Factors relating to the performance of a student, such as anxiety or examination stress that affect the student's working memory, may also interfere with the student's capacity to answer a question effectively, and may therefore cause a question to be more difficult for an individual student.

As Baird *et al.*, pointed out, 'demand and difficulty are often not distinct in students' experiences' (2009, p. 7). This study has found that students usually talk about difficulty as the catch-all term. This makes sense, from their perspective, since difficulty is what they experience in the questions. Students cannot be expected to intuit which facets of the difficulty they experienced were intended by the examiners (intended demand), and which were sources of unintended demand. Nevertheless, this study proposes that there would be a far clearer and more commonly shared understanding of the terms "demand" and "difficulty" if examiners and teachers used the two terms in a more precise manner, as has been done in this study.

In summary, anything that an examiner designs into a question should be regarded as “demand” – it is either intended demand, or it is unintended demand.

Looking at questions from the students’ perspectives, the language shifts to that of “difficulty”, either expected or unexpected. For students, there may also be a difference between “perceived difficulty” and “experienced difficulty.” Since the difficulty of an individual question, as measured and reported, is defined in terms of the marks gained by students, this is a measure of “experienced difficulty.”

Moreover, this study has also discovered, through its reflexive thematic analysis methodology and the centrality it has given to student voice, that students are able to articulate their understanding of what makes examination questions difficult. Although the extent, clarity and sophistication of their understanding varies, many students present coherent and consistent views. That many of these views buttress, mirror or extend the definitions within existent literature further points to the value and meaningful contribution of this study.

7.6 Hearing and understanding students’ voices

Students’ responses to this study’s surveys revealed that students are able to present interesting and coherent views relating to demand and difficulty in GCSE examinations, but that their voices have not previously been heard because researchers and examiners have not created structured opportunities to listen to students. In the Methods section (Chapter 4), different ways of capturing the student voice were set out, including the use of a selection of past paper examination questions, a survey involving questionnaires, semi-structured focus group discussions, some descriptive statistics, and a focus on qualitative analytical methods. In particular, and as demonstrated in Chapters 5-6, the reflexive thematic analysis technique adopted by this study yielded rich and interesting results which illuminated, through its development of themes and use of verbatim quotations, the breadth and depth of students’

understanding and views. Through its presentation and analysis of the rich and contextualised voices of students, this thesis has shown that many students are able to offer reasoned and insightful comments about the questions they encounter, and also to comment on the nature of the difficulties they face in examination questions. Those comments and responses can be constructed inductively into models of students' comprehension, and they can inform teachers' understanding of the models and misconceptions that students create.

7.7 How students experience and comprehend demand and difficulty in examination questions

The research question asked how students experience and comprehend how concepts of demand and difficulty operate in practice in GCSE mathematics questions. Through so doing, the study sought to explore the implications of students' understanding for teaching and learning purposes. The following conclusions can now be offered, from the empirical studies and reflexive thematic analysis.

1. Students identified sources of difficulty in GCSE mathematics questions that align well with taxonomies of learning and cognition, particularly Marzano's and Kendall's New Taxonomy (2007). These sources of difficulty include recall; application of knowledge; reasoning; and interpretation of information given.
2. Students associated length of question with question difficulty. This relationship appears straightforward, on the face of it: additional words increase cognitive load. This is an oversimplification, however, since a question may be made longer to support students' response strategies, or through adding a context intended to make a question less abstract.
3. Context, however, introduces an unreliable element into a question: it may be intended to lower the demand or improve accessibility, but it may, in practice, increase the difficulty

for students in unpredictable ways. Context appears to motivate some students and confuse or distract others.

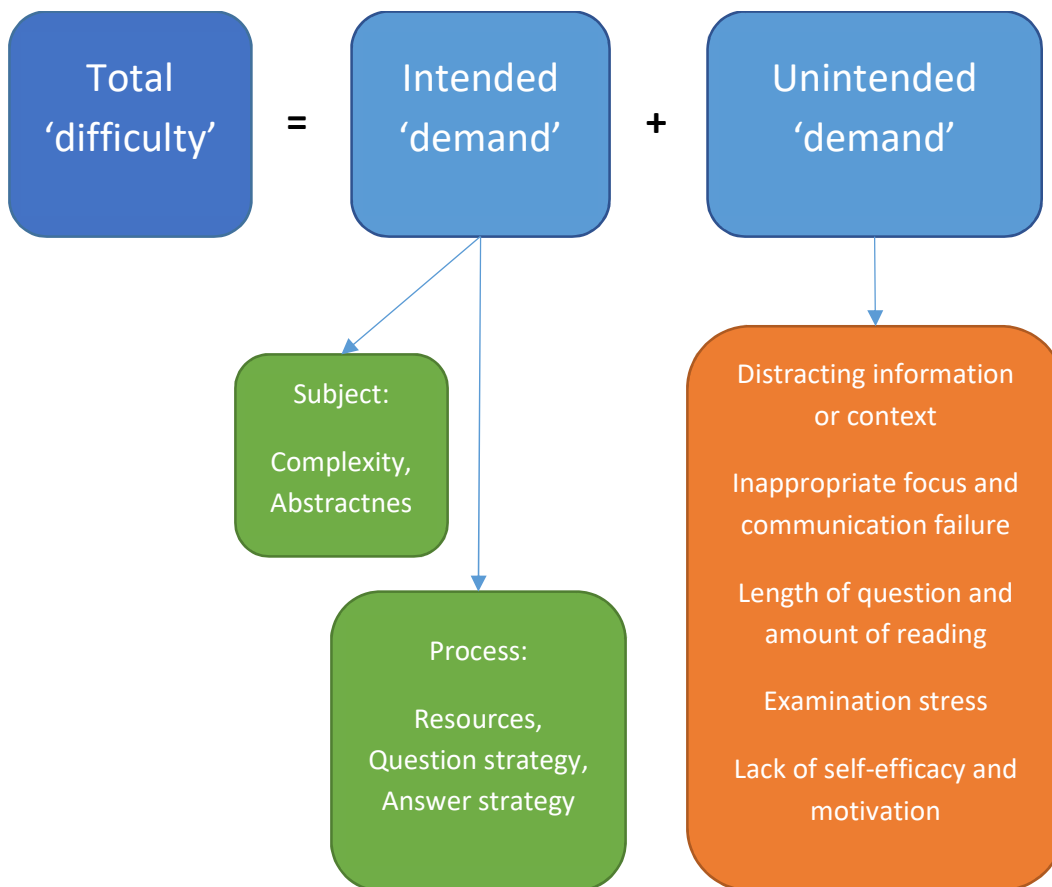
4. Students recognised that a clearly worded question gave them confidence to tackle it. Students' confidence is an important element of their self-efficacy, which in turn is positively correlated with their success (Usher and Pajares, 2009; Bandura, 1986, 1997).
5. Students recognised that there may be "distractions" from the main purpose of a question, and that these distractions have the effect of making the question more difficult. These distractions include the requirement to use "good English" or other similar wording. They may also include contextual or other features that shift the students' focus away from the examiners' main intention of the question, a conclusion from this study that further reinforces the previous findings of Ahmed and Pollitt, (2007). Distractions increase the total cognitive load by adding to the extraneous cognitive load. Examination stress can have a debilitating and negative effect on performance, reducing cognitive capacity by reducing working memory.
6. Students may not comprehend well aspects of demand intended by examiners. In particular, the requirement to select relevant facts and figures, and to select and deploy suitable methods in order to solve a problem, may be part of the design and intention of a question. Some students, however, appeared to think that this was not a fair source of difficulty, but a distraction, or even a trap.
7. Many students appeared to make rapid judgements about whether a question looked easy. Given the cognitive complexity of the process of addressing an examination question, these judgements were unlikely to be logical or reasoned; partly because thinking is slow and effortful, and partly because examinations can be high-stress environments which trigger automatic emotional and physical responses. Judgements about whether a question looked easy governed the engagement of students' self-systems. Students' ability within this study to recognise steps and apply methods, and their ensuing explanations,

can be seen to map well onto Marzano's and Kendall's New Taxonomy, in terms of their cognitive skills.

8. Students' levels of expertise, understood through analysing their responses, fit well with the Noel Burch Competency Model. Within this study, students at all stages of expertise or competence were identified, from the 'unconscious incompetent' to the 'conscious incompetent,' and from the 'conscious competent' to the 'unconscious competent.' In terms of their responses to GCSE examination questions, few students within this study fell into the category of unconscious competent. Some students gave responses that analysed into different categories for different questions, showing that their competence varied between different topics and question types.

Applying the understanding gained from the preceding survey of existent literature and analyses of the empirical studies as well as the findings from the two empirical studies which formed the primary data basis of this study, the conceptual model originally proposed in Chapter 3 can be expanded to include sources of demand and sources of difficulty. This expanded model is shown as Figure 41.

Figure 41 - Conceptual Model, with Sources of Demand and Difficulty



Source: Author's own (blue boxes based upon information from Sweller *et al.*, 1998, and Paas *et al.*, 2003; green boxes from Pollitt *et al.*, 2007; orange-red box from author's analysis).

In the green boxes on the left hand side of Figure 41, sources of demand (from the CRAS Scales of Pollitt *et al.*, 2007) can be categorised as belonging either to the subject or the process of the question. Within the subject part of demand belong issues of complexity and abstractness; within the process part belong issues of resources and strategy, either the strategies outlined in the question, or a strategy that must be constructed in the answer. To address these demands, students operate varying levels of cognitive processes (Marzano and Kendall, 2007), including first of all their self-systems (motivation and self-efficacy), and then their metacognitive systems which, in turn, regulate and control their cognitive systems. Different studies have used alternative labels for these demand features, but they agree that they are, in

some ways, desirable: Dhillon regarded them as sources of 'legitimate difficulty' (2003, p. 3); Fischer-Hoch and Hughes (1996) as 'valid' difficulty. In this thesis, they are described as sources of 'intended demand.' In terms of cognitive load theory, all of them can be classed as examples of intrinsic cognitive load. This is important to note within the context of this thesis because, according to cognitive load theory, one of the goals of instruction should be to optimise intrinsic cognitive load and minimise extraneous cognitive load. Examiners who can optimise intrinsic load in their questions will be able to create question demands that more predictably translate into difficulty for students. This might be an effective way for examiners to move assessments further along the fairness continuum (Nesbit and Shaw, 2020).

Some students understood that part of the purpose of examination questions was to test their expertise in solving problems, rather than simply to demonstrate their mathematical expertise. Understood in this way, as part of the intended demand of the questions, problem solving becomes part of the intrinsic cognitive load of the question. Many students in the main study, however, did not appear to understand this – or, at least, they wished that mathematical questions would focus only on mathematical skills and not require them to interpret contextual or embedded data – and more than one student expressed the view that such questions were '*designed to trip us up*' (R28M). For this latter group of students, demands to interpret or evaluate data appeared to be sources of extraneous cognitive load, impeding their desire simply to address the mathematical content of the question. It is interesting to note that, in the context of focus group discussions, some students appeared more willing to accept that interpreting the question was likely to be among the intended demands.

7.8 Understanding demand

Providing further depth of understanding to the research question, students in the main study identified six principal sources of difficulty in GCSE mathematics questions: the length of the question; the clarity of the wording; the complexity of the topic; the different steps involved in

forming an answer; the amount of information that needed to be recalled; and being able to identify and apply an appropriate method.

Some of these sources of difficulty – the amount of information given, the application of prior knowledge, and the need to construct a method using different steps – are evidently features of intended demand in these questions, arising from intrinsic cognitive load, and they are built intentionally into questions by examiners. Students showed that they understood how these aspects of demand worked within examination questions. These intended demand features often occur together in examination questions, as they did in the examples given to the students in this study. It was seen in Chapter 3 that the working memory model (Baddeley and Hitch, 1974) explains how new information from the question is processed in the working memory, where it can be combined with information and understanding recalled from prior learning. Since working memory is limited in capacity, the complexity of the information to be processed and the demands of the tasks in which the new and recalled information are to be used and interpreted, can easily overwhelm the student and cause cognitive overload.

Examination questions may impose high levels of cognitive load, which may be in danger of overloading the working memory. However, in written examinations, students have access to paper, which can be used to store, park and collate some of the information. They may also be able to use a calculator (in some examinations) to perform more complex calculations. When a student makes notes or sketches a diagram to assist their working out and thought processes, they are scaffolding their own problem solving, temporarily expanding their working memory in order to think through a question. More expert students will have learned to handle these additional resources more effectively. Rather as Dumbledore extracts a thought or memory and suspends it in a 'pensieve'⁸¹ in J K Rowling's Harry Potter books, enhancing his capacity to

⁸¹ "One simply siphons the excess thoughts from one's mind, pours them into the basin, and examines them at one's leisure. It becomes easier to spot patterns and links, you understand, when they are in this form." Dumbledore explaining the pensieve to Harry Potter, in Rowling, J.K. (2000): *Harry Potter and the Goblet of Fire*, Bloomsbury.

consider and ruminate on a problem, students may use paper and pen to bypass a possible 'bottleneck' (Baddeley, 2002) and extend their working memory, thus reducing their cognitive load⁸². Although students do not need literally to hold in their working memory all the information required to address a question, they will need to process this information in their working memory at some point, deciding what information is relevant, what mathematical operations to perform, and constructing a response strategy, in order to answer the question.

Some students showed their understanding that revision of prior learning helped to ease this cognitive load, by allowing the recall and the processing to take place more readily: student R27M, for example, stated that *'I think if you revise a lot then you can pick up questions a lot easier.'* Several students, however, considered that revision was *'just retrieval of given information'* (R26F) or *'just trying to remember all the correct equations'* (R63M), whereas cognitive load theory and the working memory model, building on thinking from Piaget (1953), demonstrate that knowledge is bound together as understanding in 'cognitive schemas,' achieved through continual processes of learning, rehearsal and revision. Other students' responses in the main study developed a fuller approach to the theme of memory, practice and familiarity. They explained how repeated practice helped them to remember methods and suitable approaches, and they demonstrated their confidence in handling familiar question types.

Students' partial understanding of how recall of knowledge operates within the learning process indicates that, although they evidence some understanding that relates well to cognitive load theory, they might benefit from a more in-depth understanding. The remaining sources of difficulty identified by students – clarity of wording and the length of a question –

⁸² From these insights, it is possible to train students to intentionally increase their available working memory capacity, by transferring to paper important information that is not immediately relevant. It has now become standard advice given to students about to enter an examination in the secondary school where I am headteacher, to write down as soon as possible anything that is in their head that they think they may need later, in order to allow them the working memory space to address the demands of each new examination question.

appear, on the other hand, to belong to the category of unexpected demand (the orange-red boxes on the right hand side in Figure 41), and these will be discussed later, in Section 7.9.

As discussed in Chapter 3, many teaching resources evidence only a shallow understanding of taxonomies of learning, typically going no deeper than a superficial restating of the headlines of the original or revised forms of Bloom's Taxonomy. In the discussion of the literature, it was also evident that the role of knowledge is crucial in any taxonomy. More sophisticated taxonomies of learning show knowledge being applied at different levels. Some students in this study showed that they had some understanding of this multi-layered use of knowledge: students in the main study, in particular, distinguished between simple factual recall (which they found to be less difficult) and harder factual recall. Students also made the link between the amount of practice they had had on a topic – and how recent that practice was – and the strength of their recall. These findings went against Bloom's Taxonomy's straightforward hierarchy of cognitive processes, which sites all factual recall on the same level, but it linked with Anderson, *et al.*'s (2001) revised Bloom's Taxonomy and with Marzano's and Kendall's (2007) New Taxonomy, in showing that recall can be a more or less complex cognitive process, and that harder recall imposes a higher intrinsic cognitive load. Students in the main study observed that the application of reasoning and interpreting, and the requirements to describe complex processes and give long explanations in answers, were features of the most difficult questions. It is possible to link these responses both to thinking skills higher up in the hierarchy in Bloom's taxonomy (original and revised) and the need for metacognitive skills to control cognitive processes in Marzano's and Kendall's New Taxonomy.

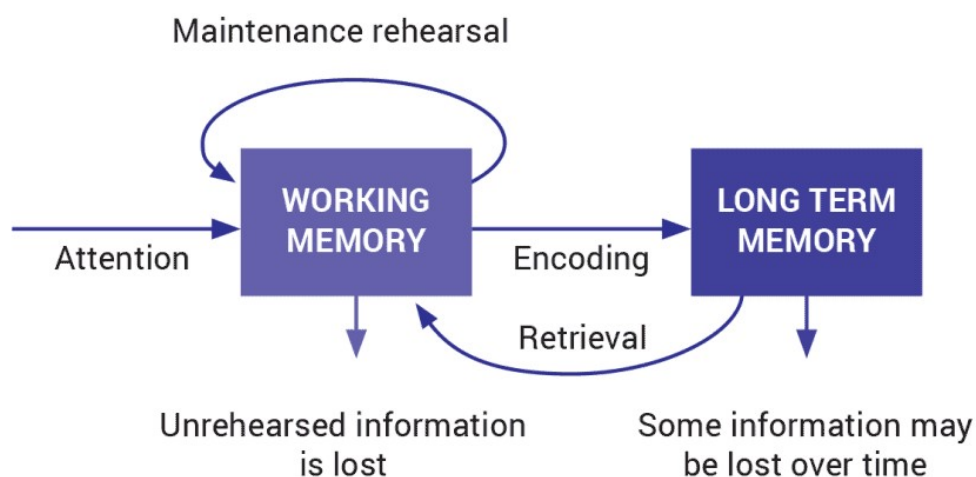
Students in the main study described the focus of their attention on first encountering a question. Whilst less confident students were uncertain where to look, more confident students had developed their own individual strategies, which they had found effective in evaluating and addressing question difficulty. Some looked to identify the question topic first; others sought out the numbers that they would be likely to manipulate; several looked for the

mark allocation, so they could understand the probable complexity of the question; a few others looked for command verbs (e.g. 'solve' or 'prove') or other indications of the question task. These are all examples of cognitive strategies being deployed by students. The Education Endowment Foundation, in its guidance report on metacognition and self-regulated learning recommends that teachers should 'explicitly teach pupils metacognitive strategies, including how to plan monitor and evaluate their learning' (EEF, 2018, p. 6). It is not evident that the focus group students in the main study are yet demonstrating metacognition⁸³, since there appears to be a lack of *intentional* and conscious selection of the most appropriate strategy with which to approach these questions from the start. The EEF notes, however, that 'it is impossible to be metacognitive without having different cognitive strategies to hand' (2018, p. 9) and, in their discussion of suitable methods to apply to individual questions, the same students evidenced more intentional choice of working methods and problem-solving strategies. Suggested steps to help students become more metacognitive include 'activating prior knowledge, leading to independent practice, before ending in structured reflection' (EEF, 2018, p. 14). Students' responses in the main study demonstrate that some of these students – and particularly those working in the top stream for mathematics – are taking these steps, leading to a growing understanding and control of their own learning. This increasingly metacognitive practice should enable them to adopt confident and effective approaches to tackling GCSE examination questions in mathematics.

Students' understanding of the strategies they employed to answer examination questions in the main study can be related to Baddeley's and Hitch's working memory model, reprised here as Figure 42.

⁸³ 'Metacognition is about the ways learners monitor and purposefully direct their learning. For example, having decided that a particular cognitive strategy for memorisation is likely to be successful, a pupil then monitors whether it has indeed been successful and then deliberately changes (or not) their memorisation method based on that evidence' (EEF, 2018, p. 9).

Figure 42 - Diagram of Human Memory



Source: Likourezes (2021).

Insights from the students in the main study show that students might begin by appraising features of the question (including topic and context, mark allocation, the words and numbers in the question, and steps required) before locating relevant information from the question, recalling prior knowledge, and then applying their thinking and reasoning skills to develop more or less structured solutions to the problem. It is fascinating to see students working through these (apparently quite makeshift) strategies in response to the demands of the question. These initial approaches, even from high-attaining students, appear less organised in the face of an examination question than might be assumed. Relating these student responses to the Baddeley and Hitch model, we can see that their disparate initial reactions are all part of the 'attention' phase – students look in a variety of places first. As they then start to engage their working memory, their attention becomes more focused on extracting information from the question, including relevant numbers and instructions. They are able to bring forward previously learned understanding and information from their long-term memory, which some found straightforward because *'we've learned it at the start of secondary and have constantly had practice by either revising it in lesson or using it to solve other questions'* (R17F), whereas others may struggle because the question is *'worded differently to how I revised it'* (R13F). Sweller *et al.*, (2011) have presented this act of recall as a complex task in itself, involving the

interrogation of long-term memory to bring forward a relevant schema and set of knowledge into the working memory.

As students combine their pre-learned and remembered methods with the new information from the question, they need to create a response strategy (Pollitt *et al.*, 2007) to the demands of the question, describing and explaining their method and deploying thinking and reasoning cognitive processes to create and present their answers. Some students appeared well aware of the possible pitfalls: *'you have to do a lot of steps and it can be easy to make a mistake'* (R30M), and more difficult questions *'need lots of different steps, and sometimes topics overlap in the steps, so it's easy to get confused/stuck'* (R40F). Other students could easily explain their multi-step method, and they appeared to breeze through the question: *'Easy, because all you had to do was take away 27,612 from 28,361 then times that answer by 0.85 then add £150 to that answer'* (Q39M). Applying these comments to the Noel Burch model, these students all appear to show conscious competence: although there are varying degrees of fluency, the students know what to do, and they apply their methods intentionally. By contrast, some other students gave responses that analyse into the consciously incompetent quadrant: *'didn't understand what I needed to do'* (Q31F); *'I think this would've involved a Venn diagram but wasn't sure how to answer it'* (Q23F).

In discussing the Baddeley/Hitch working memory model, it may be helpful to apply it to two questions from the study.

Question 6: A circle has radius 6cm. Calculate its circumference. Give your answer in centimetres, correct to 1 decimal place (3 marks) (OCR, 2018a).

In order to tackle this question, students first needed to attend to and appraise the features of the question. To take the approaches outlined by students in the main study focus groups, some might have taken in the topic (circles), others would have noticed the terms 'radius' and 'circumference,' or noted that there were 3 marks allocated, inferring that they would need to perform more than one operation to gain full marks. They then needed to recall prior

knowledge and understanding, both about the meaning of circumference and also the correct formula for calculating it. Not all students in the study were able to do this; some appeared to understand what circumference meant, but then they applied the formula for the area of a circle instead. Once students had brought their previously-learned knowledge into their working memory, they then needed to bring in the value of the circle's radius from the question (6cm), and to perform the calculation, using the value of π either in their calculator or supplied (elsewhere) in the question paper. If students had learned the formula for circumference as πd (π multiplied by the diameter), they needed to recall the knowledge that the diameter is double the radius, and therefore reason with themselves that they must double the value of the radius from the question before multiplying by π . Finally, students needed to round their answer to 1 decimal place and write it on the question paper (in this case, the survey form), also adding the unit in centimetres.

This model can also be applied to a more difficult problem, question 3 on the survey:

60% of the people in a town are males. 20% of the males are left-handed. 21.6% of all the people are left-handed. Work out the percentage of the people who are not male who are left-handed. (5 marks)

Attending to and appraising the question information, students in the main study noticed that there were several percentages within a town population, of which some were males and some were left-handed. Some noticed the instruction 'work out', implying a calculation to be performed; many others noticed the words 'not male who are left-handed'. Others again noticed the allocation of 5 marks, indicating that this was a problem with higher levels of demand. Possibly the combination of 5 marks and the wording caused some students to doubt themselves and their own skills at this point (for example, '*very wordy question which makes it more difficult to interpret. Multiple step process with use of percentages and algebra*', R36M). More successful students then recalled prior learning of conditional probability, which may have prompted them to think of structured methods such as using ratios, tree diagrams or two-way tables. Students who had not appraised the question in terms of its topic, started

some calculations with numbers from the question at this point. They also needed to construct a strategy for answering the question, reasoning that they needed to find out what proportion of the population was both male and left-handed before they could find the proportion that was not male but was left-handed. Finally, they needed to relate the proportion of the total population that was both 'not male' and 'left-handed' back to the proportion that was 'not male'. No student in the main study completed this last part of the question successfully, demonstrating that these students experienced this question as being very difficult indeed.

Using information from the question, successful students then applied their more or less structured approaches, multiplying the 60% of the town's male population by the 20% of that population who were left-handed to give 12% as the proportion of the population who were both male and left-handed. They would then be able to subtract this 12% from the 21.6% of the total left-handed population, leaving them with 9.6% of the population who were not male and who were left-handed. In the main study, 39 students out of 95 (41%) successfully applied some form of this thinking and reasoning process. Students should then have held their partial solution in their working memory – or on paper – and retrieved from their long-term memory a method to relate this 9.6% to the 40% of the total who were 'not male,' giving 24%. Few students from the pilot study (and none from the main study) did this when answering the questionnaire.

Scrutiny of the answer papers of the main study students enabled an analysis to be performed of the effectiveness of the different methods deployed (see section 6.3.2; such scrutiny was not possible in the pilot study because the survey was completed online). Students in the two focus groups spoke about their methods and strategies, and one student managed to reflect on what he should have done to complete the question; it would have been interesting to have been able to interrogate more students about their successful and unsuccessful approaches to this more difficult question. Students in the first focus group understood that

examiners may have been looking for problem solving skills, a source of 'expected difficulty' for students. However, student R31M extended this point:

'I would say – yes, they probably are looking for problem solving and that – but when they test the people and they all get the same wrong answer because they've not got the wording of the question correct, I feel like that would be because the question's worded badly. If everyone got different answers, it could be like they've all solved it slightly differently – erm, but if they get the same wrong answer – if everyone puts the percentage of the whole population instead...' (R31M).

Student R31M suggests that, if 'everyone' gets the wrong answer, being unable to cope with the additional cognitive load, the issue may lie with the wording of the question (see section 7.9.1 below).

Understanding the working memory model of students' answering processes helps teachers to see how students comprehend and experience such a question. The steps that students use and describe may be quite different from those that the teacher (and the examiner) expects. For example, a class teacher from the main study school, when asked by the researcher about question 3, stated that she would have used a ratio method, but only a small proportion of her students deployed this method in practice. This understanding may assist teachers in identifying misconceptions as students learn and also enable them to help students to take more effective control of their own cognitive processes. In other words, if teachers understand how students think and approach examination questions, they can scaffold their approach to help them to do so more effectively and consistently in future, improving the efficiency of their cognitive processes and thereby reducing overall cognitive load. Teachers would also become better at spotting students' misconceptions, which would enable them to plan sequences of learning more effectively. In the context of Question 3 in the questionnaire, the teacher would have been able to ask questions to find out where the students had become stuck or had gone wrong: it might have been in appraising information from the question, or it might have been in recalling a suitable method to apply, or in putting the information from the question into the structured method. Here, then, is a practical realisation of how the insights gained by this

study can be applied in a real life situation not only to enhance student knowledge and understanding but also develop teachers' pedagogy. In such ways, this study can be seen not only to contribute to the furtherance of existing understanding but also to improve subject-specific pedagogy and classroom practice.

Some students showed that they understood that their problem-solving techniques operated on different levels, although they did not have, within the answers given, the technical vocabulary to state this in academic terms; as illustrated by the following direct quotation from the written answer of student P48M, referring to question 1 in the questionnaire:⁸⁴

'The question isn't too difficult because of how straightforward the solution is to solve the problem, it also isn't too easy since there is significant room for error. For example, you could misread the question and multiply the 85p by 28,361 miles (instead of subtracting the mileage from when he returns the car from when he hires it), you may also mistakenly think you have to multiply the cost of £150 by the amount of miles he travels. There is enough information in the question that you need to pay attention to, but not so much that you would feel overwhelmed' (P48M).

By evaluating the student's comprehensive answer, it is possible to understand the operation of the different levels of the student's cognitive functions. First, it is apparent that his self-system is fully engaged: the question evidently interested (motivated) him, and he chose to engage with it. He also chose to set down a lengthy and interesting response, again showing his motivation; he appears to believe that he will be able to respond successfully, demonstrating high levels of self-efficacy. Secondly, the workings of the student's metacognitive system can be seen: he was able to evaluate the sources and levels of demand in the question, and he recognised potential traps and pitfalls. In fact, through explaining what not to do, or what could go wrong, the student went beyond the method, and recognised possible misconceptions. Thirdly, the student's metacognitive system was also evidently engaged with and regulated his cognitive processes in applying his expertise in arithmetic. He

⁸⁴ Question 1: Reuben hires a car. It costs £150, plus 85p for each mile he travels. When Reuben hires the car, its mileage is 27,612 miles. When Reuben returns the car, its mileage is 28,361 miles. How much did Reuben pay to hire the car? (4 marks)

saw both what and where it was necessary to subtract (mileage at the beginning from mileage at the end), when to multiply (miles travelled multiplied by 85p per mile), and when to add (the standing charge of £150, at the end) in order to supply the correct answer. To use these arithmetic functions correctly but in the wrong order, or with the wrong numbers, for example, would have produced an incorrect answer. This was an unusually complete response, from a student who was a 'conscious competent,' but many students showed elements of similar cognitive operations; thereby underlining the point that teaching that develops students' metacognitive understanding equips them very well for tackling novel problems, in examinations, in the classroom or in other contexts. This may be an example of Boud's (2000) 'sustainable assessment' – an example of an approach to formal assessment that also prepares a student for lifelong learning (see Postscript).

A number of responses showed that students had some understanding that their expertise was able to meet the demands of the examination questions, and that because of this the questions were less difficult for them. Some students in the main study felt that success comes to those who work hard. Of these, some saw the role of revision as paramount:

'I think if you revise a lot then you can pick up questions a lot easier. If you put the work in you will always get what you wanted from it' (R27M).

The meritocratic overtones here are hard to miss: according to this student, success comes to those who work. Another student concurred:

'I think the students who put their head down and are focused and determined to do well are the students who are more able as they will attempt the maths questions with full effort' (R50M).

There are references to self-efficacy here – '*determined to do well*,' and therefore presumably believing that they will succeed – but this student has conflated hard work with effort and '*ability*' (a problematic term that has been largely avoided in this study). Other students, however, saw examinations more as a memory test, with an element of chance thrown in: '*[I] just forget. Crumble under pressure. Memory test sort of'* (R41M); '*everyone makes mistakes so your result is also reliant on luck'* (R30M). Levels of self-efficacy and motivation appeared low

for these students; following the New Taxonomy, they would be unlikely to engage the self-system effectively, limiting their success.

Some students in the main study also referenced prior learning and repeated practice as factors that had enabled them to respond adequately to the demands of the examination questions. In this way, they gave supporting evidence to theoretical ideas that repeated practice transfers understanding and knowledge from the working memory to the long-term memory, from where it can be retrieved when needed.

From these students' responses it was possible, therefore, to identify not only their understanding of sources of demand, but also a number of sources of unintended demand, many of which related to concepts discussed in the literature; for instance context and focus (Dhillon, 2003; Pollitt *et al.*, 2007). In terms of cognitive load theory, these can be explained as sources of extraneous cognitive load which may take up working memory without contributing to problem solving. Some students wrote about aspects of questions that they described simply as 'confusing,' while others were more analytical.

Academic studies such as those by Dhillon (2003) and Fischer-Hoch and Hughes (1996) varied in their use of terminology, but their choice of adjectives showed that they agreed about the undesirability of these features: Dhillon (2003) described them as 'illegitimate' sources of difficulty, for example, and Fisher-Hoch and Hughes (1996) labelled some of them as 'invalid'. Aspects of a question that students find unreasonably 'confusing' are likely to reduce their perception of the question's implied contractual fairness. Identification and elimination of the effects of these sources of extraneous cognitive load would strengthen validity considerations surrounding examinations by ensuring a closer link between intended demand and difficulty. There are clear advantages to involving students in evaluating how well questions work, in this respect.

7.9 Unexpected demand and difficulty

Question features that caused unexpected difficulty can be viewed as particularly interesting within the context of this study, because the implication is that these features (and the resulting impact that they had upon the individual questions) might not have been intended by the examiners: unintended demand caused unexpected difficulty. For each type of question feature and each category of difficulty, exemplar quotations from students have been given, exploring the understanding that students bring to this important and problematic aspect of question design.

The distinction between 'intended' and 'unintended' demand is sometimes not clear-cut, however. Although many students might prefer questions in mathematics to be couched in purely abstract terms, it may well be part of the examiners' intention to test students' problem-solving capacity as well. Requirements to locate and evaluate relevant data may be part of the intended demands of the examiner, but a student might interpret these as unexpected difficulty. In the main study, student R49M stated that *'you shouldn't have to interpret anything in maths'* – for him, this additional interpretative demand was unexpected and unwelcome – while his classmate student R36M understood that *'yes, they [the examiners] probably are looking for problem solving.'*

7.9.1 Communication failure, distracting information and context

Some questions failed to communicate clearly the information and instructions that examiners presumably wished to convey. These communication failures are shown up when students' responses and answers are compared with the examiners' mark schemes and examiners' reports. Dhillon (2003) noted that communication failures caused students to make wrong assumptions about what examiners were looking for. In this study, students wrote that *'I wouldn't know where to start because it has been worded quite confusing'* (P14F); or *'I didn't understand what the question was asking'* (P42F). This study found that clearly worded questions gave students confidence, whereas questions that were unclear created confusion

and led to a loss of confidence. A threat to validity lies in that, where confusion arose, it was not possible to judge whether or not the student might have been able to address the question's demands, had they understood them. This was the case for questions 2 and 3 in the study. Such questions did not perform well as reliable indicators of the mathematical expertise of the students, and they may have led to results that misrepresented the expertise of some students.

On a similar vein of questions not clearly directing students to the problem, some questions contained elements of distracting information or context, including inappropriate scaffolding that, as Dhillon and Richardson (2003) discovered, could cause confusion. Student R31M said, of Question 3 in the main study that,

'I would say – yes, they probably are looking for problem solving and that – but when they test the people and they all get the same wrong answer because they've not got the wording of the question correct, I feel like that would be because the question's worded badly' (R31M).

Student comments that demonstrated their unease with features such as these included student P27F, who wrote in the pilot study that Question 4 was *'too complicated with too much information'*. Of the same question, another student wrote that,

'These types of questions tend to confuse a lot of people including myself because it is so much information in a small area and ultimately makes the question look harder than what it probably is' (P39F).

For this latter student, the conglomeration of *'so much information in a small area'* was confusing and overwhelming; it was a source of what Ahmed and Pollitt (2007) described as inappropriate focus, because it could cause students to mistake the point of a question or look in the wrong direction for solutions. Students identified other sources of distraction and inappropriate focus, including questions that required lengthy explanations.

Comparing two questions in the main study, a student could articulate the differences:

'For the first one, it's really clear: it says, "How much did Reuben pay to hire the car?" You know what you need to do – you just need to do the maths. But if the words are a

bit 'off', you have to interpret what it's asking before you can even start with the maths. If that's not your strong point, you've kind of failed already' (R34M).

Features of questions that distract or mislead students increase extraneous cognitive load and, hence, overall cognitive load (Lovell, 2020), making them more difficult without necessarily increasing their effectiveness as discriminators between students of differing levels of expertise or knowledge.

It is legitimate for examiners to seek to increase intrinsic cognitive load, by making students 'think hard' (Coe, 2015, p. 13), but the line between question features which do this and features which merely distract is a fine one, and a judgement call that is hard for examiners to make without understanding how students think. Student P47M illustrated this in the pilot study in his answer to Question 2⁸⁵:

'This question requires you to draw a probability tree, which can be quite tricky sometimes as you need to think about what to put on each of the branches, and usually there would be 2 different colours - not 2 different bags - which makes this question that little bit harder' (P47M).

The first part of this student's response shows his understanding of demand being created through features of the question: the intrinsic cognitive load involved in recalling a previously taught problem-solving strategy (drawing a probability tree) and applying it to the question by thinking about the labels to put on each branch. The phrase '*quite tricky*' implies that he apprehends the cognitive load involved, and thinks he will be able to manage it. The second part of his response, about the different colours of the counters and the number of bags, suggests that he is engaged by the question and regards these features as adding to the demand of the question by making it '*that little bit harder*'.

⁸⁵ Question 2: Finn has two bags of counters. He takes a counter at random from each bag. The probability that he takes a red counter from the first bag is 0.3
The probability that he takes a red counter from the second bag is 0.4
What is the probability that he takes at least one red counter? (4 marks)

The question feature of having two colours and two bags, however, distracted some students and confused others, and on balance it seems that for most students it introduced extraneous cognitive load: confusion and distraction. As was seen in Chapter 5, when responses to this question from the pilot study were analysed, there were many students who thought they knew how to do it but scored few marks. Among those who struggled, the following responses were typical: *'not sure where I would start'* (P03F); *'need more information'* (P05F); *'I think you would just have to add them, I may be wrong though'* (P14F); and *'it did not mention another colour and its a question I am not used to'* (Q47M). This question, which appeared to mislead and disconcert many students, illustrated what can happen when a question is inappropriately focused: it did not appear to direct the students' attention to the difficulties intended by the examiners. The question did not discriminate well between students of differing levels of expertise and knowledge. Interestingly, whilst only 13.3% of students in the pilot study gained full marks, in the main study 43.2% of students gained full marks, but still a fairly high proportion gained zero marks (34.7%), higher than in most other questions.

The examiners' report for this question, written after all the scripts had been marked and moderated, stated that,

'Only a small number of candidates showed an understanding of probability and offered a sound method leading to 0.58, often with the aid of a tree diagram. Very few candidates attempted to use tree diagrams and those that did so were usually incorrect' (OCR, 2018c, p. 20).

In this quotation it can be seen that examiners recognised (after the event) that this question had not performed well, but they attributed this to students' poor understanding of probability, rather than to any defect in the wording or framing of the question. Even their reference to the fact that even those candidates who did attempt to use a tree diagram to create a response *'were usually incorrect'* did not appear to cause them to doubt the merits of the question. There appeared to be a mismatch between the demands of the question, as presumably intended by the examiners (it was a question in the foundation tier paper) and the difficulties experienced in practice by the students (Chapters 5 and 6). At present, there

appears to be no systematic method to enable examiners to identify or improve their understanding of factors that lead to such ill-fitting questions: without feedback from students, examiners might well continue to believe that the low marks obtained indicated poor student understanding and technique, and they might repeat a question with similar wording in future. Teachers might also use this poorly performing past paper question for ‘mock’ exams or in teaching and assessment exercises. No information is available from OCR to confirm whether examiners discussed the performance of individual questions after the examination session was completed,⁸⁶ but it is noteworthy that no similar probability question appeared in the 2019 session, and the probability question in the 2020 examination was more conventionally worded, with different coloured counters in just one bag;⁸⁷ the examiners’ report for this question noted a much higher rate of success, stating simply that ‘many candidates were able to answer parts (a) and (c) correctly’ (OCR 2020b, p. 4). It would be in the interests of students, teachers and examiners to find systematic ways of recognising in advance, and avoiding the setting of, questions that might confuse or mislead students.

7.9.2 Length of question and the amount of reading required

As noted by Crisp and Grayson (2013), the length of the question and the amount of reading required caused students to make assumptions about levels of difficulty. In this study, it has been demonstrated that students tended to associate question length with question difficulty. Several students in the main study identified the length of question as a factor in determining question difficulty. This tendency was evident more strongly in male students:

‘The wordier a question is the harder they often are’ (R34M).

⁸⁶ The researcher’s own experience as an examiner (albeit in other subjects and qualifications) suggests that it is common practice for examiners to discuss the performance of individual questions, once the examination season is completed. To do this they would have access to students’ answer papers and to quantitative evidence of difficulty (the performance of the students), but not to qualitative feedback from students.

⁸⁷ ‘A bag contains 12 counters. 6 are red, 4 are blue and 2 are yellow. A counter is taken from the bag at random’ (OCR 2020a, p. 5). Students were asked to indicate the probability that the counter drawn was (a) red, (b) yellow, (c) green. This question was used in the main study questionnaire, as Question 7.

'When questions have more writing and are longer can also confuse the student making it more difficult' (R50M).

Students explained that short questions required simple factual recall and were not very time-consuming. This observation is in keeping with existent academic opinion; simple factual recall occupies less working memory and therefore produces a lower intrinsic cognitive load (Lovell, 2020).

Several students debated the desirability or otherwise of contextualised questions in mathematics. For some, a real-world context was a motivating factor, engaging their interest. For many more, however, context was unwelcome, seen as a complicating factor. This relates closely to research within the mathematics teaching world. Wiliam (1997) suggested, with examples, that contextual factors in a maths question can be redundant, misleading or, indeed, useful. He noted some gender differences: many female students sought to relate problems to their existing knowledge, supplying missing details from their own experience; male students, on the other hand, were often content to tackle a problem in isolation from their previous experience. In the present study, the most positive comments did come from a female student (*'I find them a lot easier, because they're more like real-life scenarios,'* R73F), but in the focus groups both male and female students preferred more abstract problems. Condensing Wiliam's (1997) research, Barton (2007) distilled three criteria for useful real-life contexts for questions, paraphrased here:

- Commonality – metaphors and contexts can be useful to aid understanding, but they must be commonly shared among all students;
- Match – the extent to which the task and its possible interpretations match the core mathematical activities intended;
- Range – how far the model takes the student along their journey to understand a topic.

Little found that many GCSE mathematics students ‘find the process of translating real-life numerical concepts into algebraic variables demanding enough, without being deflected by realistic “noise”’ (2008, p. 59). This comment finds echoes in the present study: one complained of *‘added unnecessary information that makes it appear harder than it is’* (R03F); another that *‘I do not like real-life problems and stuff like that: it just doesn’t sit well with my brain at all’* (R27M).

Little and Jones (2010) found, through experiments with post-16 students, that setting questions in a real-world context required students to interpret the question and decide what strategy to use, which imposed additional cognitive load on the student’s working memory. However, a context could, on occasion, provide useful ‘mental scaffolding’ to help the student solve the problem. They summarised the dilemma that real-world contexts therefore create:

‘On the one hand, by making a connection between the abstract world of mathematics and everyday, or scientific, contexts, we are reinforcing the utility of mathematics as a language for explaining the patterns and symmetries of the ‘real’ world. On the other hand, if we manipulate and ‘sanitise’ real-world experiences to enable them to be modelled by a pre-ordained set of mathematical techniques, then the result can appear to be artificial and contrived, or, in the words of Wiliam (1997) a ‘con’-text, providing a deception that the activity is worthwhile’ (Little and Jones, 2010, p. 138).

Barton (2007) also stated that, because of the additional cognitive load imposed by real-life contexts, care needs to be taken when interpreting students’ results. He asked three questions: whether they struggle because the fundamentals are not in place, because they do not understand what the question is asking, or because they have been misled by the context itself. Doubts around the inferences to be made from the results of a high-stakes assessment could constitute a very real threat to the validity of conclusions drawn from students’ results. In the present study, although a number of students reacted against contextualised questions, they were still able to decode the question demands and construct effective answers. It was not possible, therefore, to investigate Barton’s ideas here. Paradoxically, a few students felt that GCSE mathematics questions were not related *enough* to the ‘real world.’ One complained about *‘wording questions in ways you would never actually see in the real world’*

(R63M) and another observed drily that the *'real world is simply different to questions on paper'* (R56M).

Longer questions are often broken into steps, or sub-questions, or require the construction of an answering strategy that proceeds by step. A theme of "recognising steps" was identified in the pilot study, and a theme of "steps and methods" was developed in the main study.

Students in the pilot study tended to state they were confused by the more difficult questions, but in some cases the levels of engagement or resilience appeared to have been relatively low (an evaluation based on the short amount of time students spent on the online questionnaire).

Some students, in both parts of the study, correlated the number of steps required in a questions with the level of difficulty (more steps = more difficult). This common student response, however, again demonstrates the complexity of applying design features to questions intended to modify their demands. Whereas, for most students, a question with more steps looked more difficult, some examiners might have intended to break down the question into different steps in order to improve accessibility, as discussed in Section 2.1.

Some higher attaining students were able to articulate complete methods, demonstrating their comprehension of the different steps required by the question. A few of these students were able to relate their understanding of the method to their evaluation of the difficulty of the question. These students showed higher levels of metacognitive understanding.

Examiners ought to (but possibly do not) appreciate that, to many students, a multi-step question can appear daunting, implying a complexity and difficulty that may not be borne out in practice. Since students' self-efficacy and resilience may not be strong, this could cause students not to attempt questions they could successfully answer if they appeared as individual questions, perhaps leading to an inaccurate reading of their actual expertise. Some students appeared to grasp, at least in retrospect, that length does not necessarily equate with complexity, but many did not. Examiners who understood this could carefully plan the format and appearance of questions to enable students to approach them confidently; and teachers

who understood this could help build practice with students from an early stage to accustom them to multi-stage questions.

There is a validity argument to be made in the opposite direction, however, against the oversimplification of questions. Many employers report that they want students to have solid problem-solving skills⁸⁸. Such employers might expect to interpret a student's high grade in GCSE mathematics as an indicator of problem-solving skills. The validity of such an interpretation would be undermined if mathematical questions did not require students to decode instructions, select and evaluate data, and construct/apply appropriate methods. Some students understood this: student R30M recognised that one of the aims of a mathematics examination question was that it *'rewards problem solving and creative thinking.'*

7.9.3 Examination stress

Examination stress and performance anxiety are 'negatively related to examination performance' (Putwain and Symes, 2018, p. 482). Some students in this study recognised the debilitating effect of pressure and stress, and the possible consequences of poor performance.

Students wrote:

'I think high-stakes questions like 6 marks are more difficult because rather than just being able to do the maths if you don't know the topic it freaks you out because it can be a difference of a whole grade' (R31M).

'Some people struggle with the pressure of tests and don't achieve their full potential. I don't think it is fair on them. Also everyone makes mistakes so your result is also reliant on luck' (R30M).

'I personally find it a lot of pressure in exams and forget a lot. I also think some people run out of time and a lot of things can go wrong so I think it is an unfair way to compare' (R29F).

⁸⁸ According to the Chartered Management Institute, in a survey of employers, critical thinking and problem-solving skills were ranked second highest, behind teamwork skills (September 2021): <https://www.managers.org.uk/knowledge-and-insights/article/the-skills-that-employers-want-in-the-modern-workplace/> accessed 8.5.2023. In the Job Outlook Survey (2022) conducted by the National Association of Colleges and Employers in the USA, nearly 86% of employers said they wanted proof that prospective employees have solid problem-solving abilities.

Some students referenced examination stress and performance anxiety as factors which had affected their capacity to answer questions well. Recent studies have explored the causes and effects of examination-related stress among students taking GCSE examinations. Putwain and Aveyard found that students with higher levels of perceived control over their outcomes performed better at low levels of worry; as examination worry increased, 'the differential advantage offered by higher perceived control diminished' (2018, p. 65), to the point where, at high levels of worry, control made little difference to performance. Putwain *et al.*, (2012) suggested that a small degree of examination stress could actually improve performance, through increased motivation to a given task that led to greater efforts being made by the student; but Putwain and Symes (2018) found that, even with increased effort, the combination of high cognitive load and examination stress could often overwhelm students. Students in this study, who expressed a view about examination stress, saw it as having a negative, debilitating effect, which accords with the majority of existent research.

Beilock (2008) comments that cognitive capacity is impaired by stress, and that this can result in a reduction in working memory capacity being available for the cognitive tasks of an examination. This is not restricted to UK students: Lau *et al.*, (2022), using data from three large-scale international assessments of student achievement in mathematics, suggested that individual students' anxiety in maths was negatively correlated with maths achievement across the globe. It is clear, from research studies and from the verbatim comments of the students in this study, that many students approach examinations – in mathematics in particular – with high levels of anxiety, worried about their possible performance, the expectations of others, and the consequences for their future prospects. One even described a question as '*intimidating*' (R34M). Any question features that create unexpected demand and impact on the capacity of students to demonstrate their expertise are therefore likely to have an increased distorting effect in the context of mathematics examinations, and it is important that examiners can both detect and avoid such features in their questions, in order to report accurate readings of the expertise and knowledge of students.

7.10 Self-efficacy, student motivation, and fear

Self-efficacy – a student’s belief that they can be successful when carrying out a particular task⁸⁹ – has been shown to be a predicting factor in students’ capacity to meet the demands of learning activities and test questions (Bandura, 1986). A lack of self-efficacy may afflict a student at any stage in the problem-solving process: it may prevent the self-system from engaging at all, or it may inhibit working memory capacity. A student’s confidence and self-efficacy may be quite fragile, and can be easily eroded by failure or the look of something unfamiliar. These are insights that were generated during the analysis of students’ responses.

Bandura’s psychological assessments had led him to assert that,

‘Students who develop a strong sense of self-efficacy are well-equipped to educate themselves when they have to rely on their own initiative’ (Bandura, 1986, p. 417).

Self-efficacy may, therefore, have a positive influence on learners’ capacity to tackle unfamiliar tasks in examination situations. In this way, self-efficacy is the psychological basis for the confidence evidenced by some students in the empirical parts of this study (Chapters 5 and 6). Teaching strategies that promote and help students to develop self-efficacy are therefore likely to have positive effects in helping students prepare for unexpected question features in examinations. Training in such techniques within the classroom may include appropriate scaffolding in earlier stages of instruction, thus building confidence, and the regular exposure to questions carefully chosen for each student by their teacher so that their demands are located within the student’s Zone of Proximal Development (see Section 3.1.3). It might also be advantageous for all teachers and examiners also to understand and apply the principles of cognitive load theory, making learning more effective and examinations more reliable by minimising extraneous load and optimising intrinsic load. However, the reality may be quite different, and it may be impossible in practice for examiners to predict the expertise and prior

⁸⁹ Cambridge Dictionary, online, accessed 12.07.2023

learning of students, or the ways in which features of their questions may affect individuals. It would therefore be desirable if students could be taught to apply cognitive load theory principles themselves to manage their own cognitive load and working memory, in the classroom and in examinations. Indeed, Sweller *et al.*, (2019) suggested that students who were taught to understand and apply cognitive load theory principles might be better equipped to manage their own cognitive load in the face of unexpected and poorly constructed learning (or test) situations.

Applying these insights from literature to the present studies, student P30F's self-efficacy and confidence appeared high: she found the question *'easy, because I understand the question and will be able to give a good answer'*. Student R08F showed lower levels of self-confidence: *'I wasn't fully sure how to start working it out, and I'm not sure I did get it right.'* Student Q23F, by contrast, started off confidently, but her resilience crumbled on contact with a more demanding question: *'understood both [parts] a and b, confident in solving these equations. c) have no idea where to start the question.'* Students often appeared to make rapid judgements about whether a question looks easy. There is a link with students' self-efficacy here. Semantic indicators in students' explanations (such as the use of the word 'just') showed levels of confidence. Some students were right to be confident in their own performance: for them, the question "looked easy – and it was." For other students, their confidence was misplaced: for them, the question "looked easy – but it wasn't." Although a large number of other studies – see, for example, Pajares and Kranzler (1995); Erickson and Heit (2015) – suggested that a tendency towards over-confidence might be more prevalent among male students, particularly with regards to mathematics, this study found the tendency to be spread almost evenly between male and female students, albeit with small numbers self-reporting. It is beyond the scope of this study to offer possible reasons for this difference, but it provides a further possible avenue for future research.

In the main study, fear and enjoyment were developed as aspects of the “motivation” theme. Some students found questions difficult, and that could be demotivating. Jackson (2010) found that students’ behaviour was often motivated by fear, particularly of academic failure, and that ‘these fears are particularly pronounced in relation to exams, and especially those that are used to rank schools publicly’ such as GCSEs (Jackson, 2010, p. 190). In a competitive school system, such as the one in which schools in England operate, Jackson states (2010, p. 190) that ‘the stakes are high so there is considerable pressure to be a winner, and possibly even more not to be a ‘loser’.’ Alongside fear of academic failure, students also fear social failure, fear of not fitting in. For some students, this can be motivating to work harder, and this is a theme within some of the responses in this study. As Jackson notes (2010, p. 195), the ‘dominant pupil discourse within many secondary schools that it is uncool to work hard’ and those who are academically successful can be labelled as ‘boffins’ or ‘geeks’ (Francis *et al.*, 2010). For students, including those who have been less academically successful during their time at school, there can be a strong social reaction against working hard: in their eyes, they have already failed academically, and it would be ‘uncool’ to try: these students ‘fit in’ by not working hard. Thus, Jackson argues (2010, 2017), fear of social failure may actually contribute further to academic failure; it may contribute to mental ill health, particularly for female students (Stentiford *et al.*, 2021).

Just as fear of social failure might motivate some students to turn away from academic effort, however, in the present study it was evident that social success was a positive motivation towards seeking out more difficult questions. First, students spoke of their enjoyment of challenge, and that they liked being able to succeed in tackling difficult questions that others found too hard: *‘I just like always being challenged by it’* (R73F);

‘When they’re really, really hard and not many people get it, there’s a select few... who will get it first try. And I prefer it when it’s like that’ (R77M, fillers and repetitions removed).

Secondly, students enjoyed the social side of academic success:

'I think sometimes it does help to, like, explain things to other people... if we're going through a question and I sit next to my friend and we're going through a question it, like, helps me understand... cos it shows we both know something about it' (R73F).

Student R77M, having explained above that he enjoyed getting the really hard questions right, went on to explain that this also had a social value for him:

[Student R87M] 'quite often asks us for help. I enjoy that – I actually enjoy helping people: that's fun for me' (R77M).

Student R34M contextualised this social love of learning, showing his enjoyment of the company of others who were also high attainers, but who had different approaches from him to particular questions:

'So there might be like question 4 – I could do something completely different to [student R36M] and get the same answer and it's just about working and finding a way that works for you to answer the question. Cos [student R31M] does crazy little equation things for stuff and it just doesn't make any sense at all' [laughs].

[Researcher:] Doesn't make any sense to you?

'Yes! But it obviously like makes sense to him, cos he's a genius child' [laughter] (R34M).

The social cohesion and mutual motivation of these high attaining male students shines through their words and interactions with one another. For them, the challenge and opportunities for competition that demanding questions bring is welcome and enjoyable; their ability to tackle hard questions provides opportunities for social bonding.

Through the themes discussed in depth in this chapter, this study demonstrates that students exhibit levels of understanding around their cognitive processes that relate well to a number of learning theories. In this way, students' responses give practical confirmation to these learning theories. In general, students understood well the link between the demand of questions and the performance of students.

Students in the pilot study were more accurate in their estimations of the difficulty of questions they had already attempted than of questions that they had merely looked at. It is perhaps not surprising that students appeared to find it harder to make absolute judgements of demand and difficulty when they had not already attempted to answer the questions. Students in the main study found it straightforward to make comparative judgements of difficulty between pairs of questions on the same topic that they had already answered. This corresponds with the findings and observations of Christodoulou (2016), and with the work of, amongst others, Pollitt (2012), and Jones *et al.*, (2015) with regard to using comparative judgement techniques when assessing students' answers to examination questions. The findings of this study, therefore, can be seen to further buttress and support these elements of existent literature.

The importance of GCSE examination questions in producing a fair and accurate reading of a student's expertise and knowledge was mentioned by a few students; this was because of the role that their examination results can play in determining their future. This consequence-laden view of examinations broadly correlates with Kane's (2016) interpretative arguments pertaining to validity and the purposes to which the results of a test are applied that validate the test. Once more, therefore, the views of students expressed in this study can be seen to further confirm the findings of existent literature whilst also providing additional valuable contributions to the knowledge and understanding of how – and how well – the GCSE examination system really works.

7.11 Comparison of examiners' and students' views

The findings of this study suggest that examiners have views of what creates demand in examination questions that may not align with what students experience and what happens in practice. Examiners may, for example, believe that a real-life context can help a student to relate to an examination question whereas – and as was seen in this study – some students may find such contextualisation to be confusing or distracting. This mismatch between

examiners' anticipation and students' reality may lead to the creation of examination questions that perform in ways that are different to those expected and predicted by examiners. In the absence of the pre-testing of examination questions in the UK,⁹⁰ there is at present no pre-examination mechanism for adjusting or weeding out questions that do not perform well. Although pre-testing is unlikely to become a widely-used evaluative tool in UK examinations, owing to financial and logistical constraints, it would be possible to combine a thorough post-test analysis of the performance of individual questions – perhaps using Rasch analysis – with the pre-testing of a range of possible questions, possibly using a focus group of students so as to make the exercise manageable and affordable. This would give examiners access to a rich data set of evaluative materials that could be used to improve the wording and performance of examination questions.

It has also been argued, that the wording of an examination question should focus the attention of students towards the intended problem:

'Question writers should want to cause all of the students to be trying to answer the intended question: only then can we assess how well they are doing it' (Ahmed and Pollitt, 2007, p. 206).

A question such as Question 2 in the study (the probability of taking a coloured counter from a bag), which many students felt looked straightforward but which they then found that they could not answer, and which was misinterpreted by many others, may be an example of a poorly focused question. Some students were particularly aware of the confusion caused by presentational factors in the question. Given such weaknesses, this question could not, therefore, give results that could be interpreted validly as a measure of students' expertise. It follows, all else being equal, that according to Kane (2012), this would be a matter of construct

⁹⁰ The reasons for not pre-testing questions that were put forward by Baird and Black (2013) include that pre-testing is extremely expensive, so it cannot be justified by examination boards for questions that are released into the public domain once the examinations have been sat, meaning that questions cannot be banked for future use. Specifications (syllabuses) also change frequently, so questions could become out of date quite soon. In short, the educational and validation benefits of pre-testing items are outweighed commercially by the costs.

validity: results of the examination (and, therefore, its component questions) may be used to infer judgements about the expertise and knowledge of students, and their suitability for further study or employment; if the examination (or question) does not perform as expected, the consequential inference from its results may not be valid. A typical GCSE examination paper in mathematics might contain 20 questions, so the distorting effect of one question is likely to be diluted, but the 4-5 marks available from an individual question might still make the difference between one grade and another for a student.

With regard to precisely predicting the difficulty of individual questions, it remains the case, as Baird and Black commented, that even expert examiners cannot always precisely predict the experienced difficulty of an individual question and that, as a result, there are always 'surprises in examination results about which questions were most difficult' (2013, p. 11). This view was evidenced within this study; a few students doubted the intentions of their examiners, and they were also clear about the effects of misleading questions: student R31M, for example, railed against *'misleading'* words that cause *'misreading or even misinterpreting questions. Minor mistakes that end up costing the whole question.'* Student R28M thought, *'they are designed to trip us up.'* It is the contention of this study that a closer link between the theory and practice of assessment would be highly beneficial.

Such confusion may lead to student under-performance. Within this study, the responses of students illustrate that a number understood and were anxious about the consequences of under-performance in high-stakes assessments such as GCSEs owing to factors beyond their control. Student R31M wrote that *'tests as a whole can be stressful or give students an unfortunate set of questions that they may have otherwise been able to get correct;'* student R60M said, *'you are under more pressure due to the fact one question can impact your mark largely'* and added, *'but then it's these tests which may determine your future.'* Such responses further underline the appropriateness of this study's contention that there is a need for students' perspectives to be included in evaluations of examinations, so that examiners have a

good understanding not only of the performance factors of individual questions but also of the construct validity of the assessment process they control.

Having presented and discussed many different aspects of students' experiences and understanding, it is now time to offer some conclusions and recommendations.

This page is left intentionally blank

Chapter 8: Conclusion – Reflections on the journey

In this chapter I make some final reflections on my journey, in the form of conclusions, together with some insights and “suggestions for future travellers”. Like T.S.Eliot’s magi, I find myself ‘no longer at ease here, in the old dispensation’⁹¹. I offer recommendations and opportunities for further research.

The overall research aim of this study was to investigate the “views from the students’ desks” so that greater understanding could be achieved with regard to what is understood by the concepts of “demand” and “difficulty” within an examination context and how these work in practice. In addition to commenting on these two concepts and directly addressing the research question of this thesis, this concluding chapter also notes the limitations of the current study, makes practical recommendations and applications of the findings of the research, and proposes a series of future research avenues which build upon the findings of this study.

8.1 A better understanding of what is meant by “demand” and “difficulty”

With regard to the two primary concepts that have underpinned this work, this thesis has revealed that, presently, there is no single, clearly defined view of what professional educators understand by demand and difficulty. Some use the words interchangeably, and others appear to have an element of confusion about the distinctions between the two terms. Only a few researchers, notably those who have been immersed in both academic research and examination board practice – such as Jo-Anne Baird and Alastair Pollitt – have been clear about the meaning and use of these key terms. This thesis widens existent knowledge and understanding by clarifying the definitions of these two terms and relating them, via a conceptual model, to a theoretical framework of demand and difficulty in public examinations.

⁹¹ T.S. Eliot – *Journey of the Magi*, 1927.

Demand, this thesis contends, is a multifaceted concept. Within the context of examination papers it relates to the aspects of an examination question that create intrinsic cognitive load, and it should be within the control of examiners who create the questions. Moreover, factors that are involved in creating demand in examination questions include, as Pollitt *et al.*, suggested in their CRAS scales (2007), aspects related to the subject or topic, and aspects of the processes required of students in answering the questions. Allied to these aspects of demand, subject-related aspects of the concept pertain to the complexity and abstractness of topics, whereas process-related aspects of demand are the resources needed to answer the question and the strategies required, both those outlined in the question and those required of the student in forming an answer. These varied aspects of demand impose intrinsic cognitive load on students answering questions. Research into cognitive load divides the load delivered by the demands of a task or question into intrinsic and extraneous cognitive load; it is considered desirable to maximise intrinsic load and minimise extraneous load. As has been seen in the Discussion chapter, there can be some debate about whether particular features of a question are considered to be part of intrinsic or extraneous load. Given this, it might be thought desirable that examiners who set examination questions should understand cognitive load theory sufficiently so they can manipulate demand factors to create the desired level of cognitive load in their questions and examination papers. It has been shown that whilst demand is qualitative in nature, it can also be evaluated, graduated and compared. Although the demands of a particular question are fixed, different students may experience these demands differently, depending on their levels of expertise and their particular approach.

In terms of defining and understanding difficulty, the concept is more straightforwardly quantitative in nature. As this study has illustrated, the difficulty of a question is measured by the proportion of students who are able to answer a given question correctly. The difficulty of a question is, therefore, inversely related to the average mark on that question gained by students: a low average mark connotes a high level of difficulty, and *vice versa*. Increasing students' expertise causes them to answer more questions correctly, and therefore to

experience these questions as being less difficult. Aspects of demand ought to be the principal cause of difficulty experienced by students. Unexpected demands are likely to translate into unpredictable difficulty.

8.2 Rebalancing the relationship between examiners' and students' perspectives

Closed book invigilated examinations dominate the western Anglophone examination and assessment system, and it is within this context and within English schools in particular that this study has been conducted. In the Introduction, a simplified model of the examination system was set out under the guise of the Utopia Examination System Ltd. Anatomising this system led to speculation about the processes that might occur inside the examination system. Having acknowledged, however, that this simplified model does not exist (it never did, it could be argued, and it is unattainable), any pretence to a positivist ontology was willingly abandoned: there was no objective hypothesis to test, and there could be no objective research tool to examine it. Just as an idealised 'noise-less' examination cannot exist, it follows that a positivist view of examinations also cannot logically stand up. A positivist view of an examination would be that an objective reality – in this case, an objective measure of the student's true expertise – exists and can be found. On the contrary, this study has established that any claims to knowledge are, in keeping with the views of Fosnot (2013), emergent, developmental, and non-objective. This perspective applies not only to the understanding that has been developed here through investigation of students' views but arguably also to the very measures and judgements that the examination system seeks to make about the knowledge or expertise of students. Just as the constructivist ontology adopted in this study has given rise to an interpretative epistemology, an imperfect and socially-constructed examination system cannot help but deliver results that are, at best, merely indications (rather than empirical measurements) of a student's attainment and expertise. In this light, examination results should be seen as examples of 'accounts and observations of the world that provide indirect indications of phenomena' (Waring, 2021a, 16).

Such a conclusion, rooted in both established academic opinion and the specific results of this study, is not intended to undermine confidence in the examination system, however. Rather, this position signals two constructive developments, representing a widening of existing thought and understanding of processes inspired by this study. The first is that there should be continual efforts to improve examinations systems: the fact that examination results give an imperfect reading of a student's knowledge and expertise does not signify that the result is without meaning. Rather, it indicates that the meaning is more complex and nuanced than most people might conveniently prefer to think. Nevertheless, efforts ought to be made to make the meaning and its consequent interpretations as reliable and valid as possible.

Secondly, an examination result ought to be interpreted as having meaning in its own context: those who wish to use or interpret an examination result should ask, "what grade?", "which topics were covered and at what depth?", "what skills and knowledge were tested?", "when was it studied?", and so on.⁹² In the context of relatively stable grading proportions and grading criteria, a particular grade at GCSE may be thought to convey the attainment of a certain standard within a subject. However, the added knowledge provided by this study is that, if a student underachieved in one area, not because of their lack of expertise but because of additional difficulty introduced by unpredictable question features, the examination has no mechanism to compensate for this; the student's individual grade is still judged against the performance of other students, and against the grading criteria set by the examination board. There may therefore be an impact on the perceived relational fairness of this test.

This thinking ought now, this thesis contends, to be extended so that an understanding is established that an examination grade is but one indicator among many of the skills, knowledge and expertise of a student that may be garnered. It follows, in such a more nuanced world, that to make valid inferences from an examination grade, it may be necessary

⁹² Some application processes already require additional information such as this, for example from university graduates about their degree courses.

to engage in processes of further validation, and to seek triangulation from alternative sources of evidence of knowledge and understanding. This is a process that would, building on the findings of this thesis, offer future opportunities to reinvigorate the debate, between examiners, researchers and with the general public, about the purpose and function of public examinations such as the GCSE within society.

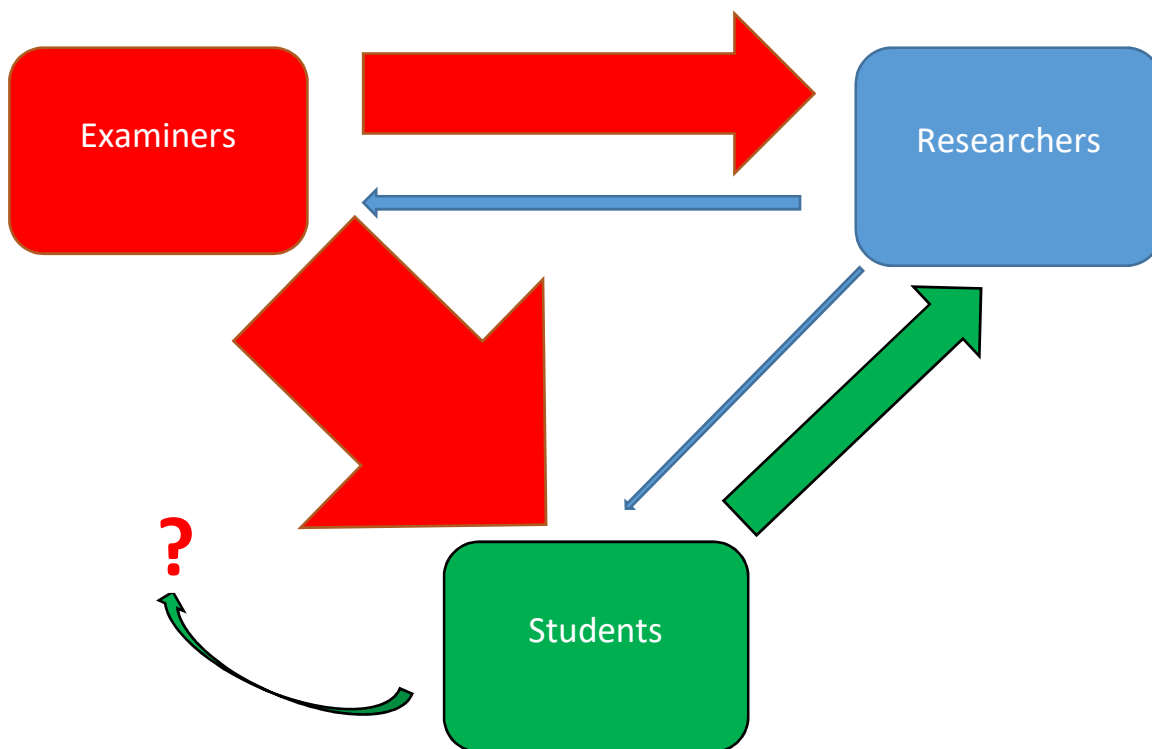
Some authors have urged examination boards to take more account of students' views. Carless (2009) asserted that students need to assume a 'mediating function' in negotiating matters relating to the fairness of their assessments. Building on the suggestions of Entwistle (1991), Orr (2010), McArthur and Huxham (2011), Nisbet and Shaw stated that,

'Those responsible for assessments should... have a means of recording any suggestions about how the work that they are doing many contribute to greater fairness and any concerns about foreseeable unfair uses' (Nisbet and Shaw, 2020, p. 160).

From a general societal perspective it might be assumed that the relationship that exists between examiners, researchers and students might simply be a reciprocal triangular one, where each of the three parties contributed to and gained from each of the others in roughly equal measure.⁹³ This study has revealed, however, having looked in more detail at how the relationship works in practice, that a more complex picture needs to be constructed. Figure 43 (overleaf) illustrates the relationship between examiners, researchers and students, as the findings of this study suggest it actually is. The relative size of the arrows gives an impression of the relative influence of one party on another. This relationship appears, from the perspective of this author, fundamentally unbalanced.

⁹³ For example, Puttick (2015) found that some teachers assumed as a matter of course that chief examiners, as subject specialists, were also researchers.

Figure 43 - Relationship between Examiners, Researchers and Students



Source: Author's own.

The largest arrow indicates the influence that examiners have on students. Examiners (using this term to denote those in control of examinations, who set questions, and who oversee marking and awarding processes) create specifications for examinations, determining the knowledge that will be assessed; they create examination questions; and they put questions together to make examination papers. Thereafter, and after the actual sitting (by students) of the examinations, examiners oversee the marking, standardisation and moderation of examination papers; and they publish the results; the latter, as noted, being a process that determines the future directions and opportunities open to students (both individually and collectively). Finally, and following the publication of the results, the examiners may publish a report, in which they may describe and reflect on the performance of the examination questions that they (and others in their team) set, marked, and moderated. No one examiner does all this, of course, but it is in the control of the examination boards.

Beyond this already substantial influence, examiners' influence on students may also be seen to extend well beyond the examinations themselves. There is very substantial 'washback' from assessments – particularly, public examinations such as the GCSE – into the school curriculum (Alderson and Wall, 1993). Some teachers also work as examination markers, often so they can gain a better understanding of how the examination system works in order to inform their own teaching, strengthening this 'washback' effect. Puttick suggested that examiners assume the roles of 'both Prophet and Priest, being actively involved in the construction of distributive, recontextualising, and evaluative rules' (2015, p. 481). This almost mystical, absolute power over the curriculum means that both what is assessed and how it is to be assessed dictates and dominates what is taught and how it is taught.⁹⁴ McEwen made this link succinctly: 'what is assessed becomes what is valued, which becomes what is taught' (McEwen, 1995, p. 42), whilst, as Baird *et al.*, observed, 'the taught curriculum was narrowed to the material that was anticipated on the test' (2017, p. 319). In this way, high-stakes testing drives teachers and students to change their behaviours (Stobart and Eggen, 2012). Examiners therefore dictate not only what knowledge students will learn, but also what knowledge is valued by the educational community at large. Ultimately, therefore, they shape and determine what knowledge is passed down to the next generation. It is a philosophical and moral question as to whether examination boards, which are accountable only to a government-appointed watchdog, should be the arbiters of knowledge in society in this powerful way.

Within what might be viewed as an overly examiner-centric system, there is no mechanism or forum for the views or values of students to be heard by (much less to influence) examiners. This represents a power imbalance and is shown within Figure 43 by the fact that the return arrow between students and examiners is shown as not connecting them but ending in a

⁹⁴ Puttick (2015) gives an example of a geography teacher who felt the need to apologise to her class for having inadvertently spent a lesson investigating a case study that was no longer on the GCSE specification.

question mark.⁹⁵ To “correct” this power imbalance, or at least to start to address it, this thesis argues – as discussed further in Section 8.4 (Recommendations) – that it would benefit examiners to hear and evaluate the voices of students, so that they may gain valuable information about the performance and effect of their questions, and a broader perspective of the experience of the students who have the most to gain and lose by taking their examinations.

The performance and nature of examinations are evaluated and investigated by researchers, who publish reports and articles. Some researchers work as part of the teams for examination boards, and their areas of study and publications are highly relevant to the work that examination boards carry out. Nonetheless, it has not been possible to establish, in the course of this study, that their research has influenced the practices and thinking of examiners in any substantial way. References to research were at best superficial in the information available to prospective examiners – including details of training for examiners – published by AQA, Edexcel/Pearsons and OCR⁹⁶. The arrow pointing from researchers to examiners is therefore narrow.

Researchers also study the work and output of students. Their publications may influence the training, development and work of teachers and, through the teachers, the ways in which students are taught. Because the work of examiners and the work of students form the basis of much of the work of educational researchers, there are substantial-sized arrows from these two groups towards researchers. There is some influence from researchers towards teachers and, via teachers, to students, hence the thinner arrow from researchers towards students.

⁹⁵ An exception is possibly that if too few students choose to study a subject, it may become commercially unviable for an examination board to continue to offer the subject, but these are arguably market forces, potentially subject to educational fashions and political pressure, rather than the wills and opinions of students. See, for example, the dramatic drop in entry numbers in GCSE Religious Studies after the subject was removed from the ‘Humanities’ section of the so-called ‘English Baccalaureate’ in 2011 (*Guardian*, 2011), and the sharp decline in numbers taking GCSE and A level Music between 2008 and 2019, when arts subjects were no longer given equal value in government metrics for school performance (ISM, 2019)

⁹⁶ Examination board websites, accessed 07.02.2022

Students study the specifications for examinations published by the examination boards.

Examination questions dominate the final years of secondary education in the UK: students practise answering examination questions in class; they sit 'mock' GCSE papers; they receive feedback from their teachers on their performance in these practice questions, and they may use past paper questions in their revision and preparation for public examinations. After these GCSE examinations have been sat, students receive marks and grades. Students' future progress often depends on these grade outcomes: the grades may determine whether they go on to study further and, if so, which courses and which subjects they pursue. Their GCSE examination grades may well determine their future career and earnings. The effect of the work of examiners on the lives of young people would therefore be hard to over-state. By far the broadest arrow in Figure 43 therefore points from examiners towards students. As has been seen in this study, there is no formal mechanism for students to give any feedback to examiners.

This lack of balance represents a lack of equity in the system. Since any unfairness in educational systems affects disproportionately those who are already disadvantaged in society,⁹⁷ this lack of balance is also a matter of social justice. The effects of the unbalanced relationship between examiners and students are that examiners do not know how effective their questions and papers are in accurately measuring the knowledge or expertise of students. They do not know whether the demands they intend in their assessments carry through to become fair and equitable sources of difficulty for students, and therefore they do not know whether the GCSE grades they award can be said to be valid and accurate representations of the knowledge and expertise of the students. Examiners have no control over the interpretations and inferences that others in society (such as employers and course admission tutors) will make from the grades output from GCSE examinations. It would therefore be reasonable to suggest that they should seek to use all tools at their disposal to

⁹⁷ Institute for Fiscal Studies: The IFS Deaton Review, August 2022:
<https://ifs.org.uk/inequality/education-inequalities/> accessed 29.05.2023

understand and improve the ways in which their intended demands in examination questions translate into experienced difficulty for students. Listening to the voices of students in a structured way might well help them considerably.

8.3 Hearing students' voices and investigating their understanding of demand and difficulty in examination questions

Methods deployed over the course of this study, both within the pilot study and the main study, created opportunities for students to engage with actual past paper examination questions and to give their considered responses to detailed enquiries about their reactions. A process of reflexive thematic analysis was then deployed to engage with, interpret and develop meaning from these responses. Reflexive thematic analysis is a robust and accessible method that analyses, develops and interprets patterns across a qualitative dataset. It allows the voices of students to be heard and attaches value to what they say; as such it could be used more widely by teachers and researchers.

This study's research question focused on how students experience and comprehend how concepts of demand and difficulty operate in GCSE mathematics examination questions, with the aim of presenting these experiences and this comprehension and to probe the implications for learning. The methods for the study were outlined in Chapter 4; results were presented in Chapters 5 and 6, and they were discussed in Chapter 7. The pilot study provided invaluable information and responses from students. A thorough evaluation of the pilot study led to the design of the main study. This produced a wealth of detailed student voice commentary that was analysed, interpreted and developed to create meaning around students' varied experiences and comprehension of demand and difficulty in GCSE mathematics examination questions. Students' insights extended beyond matters of learning and cognition, to other issues of concern and interest around high-stakes examinations. These included reflections on

examination stress, the value of revision and hard work, students' motivation and the social benefits of academic challenge, and the consequential validity of examination grades.

This study therefore demonstrated that the responses of students can be collected, analysed and distilled, and that qualitative methods can make a valuable contribution to the academic and professional discourse on teaching, learning and assessment. Students, it was revealed, did not distinguish between concepts of demand and difficulty, since all sources of challenge – whether intended or not, and whether they represent intrinsic or extraneous cognitive load – appeared to them as sources of difficulty. They were not able to discern which elements of this difficulty may have been intended by examiners and which not, but they did sense whether the difficulties they experienced appeared to them to be fair and reasonable, or whether they were “tricky” and distracting.

The principal method for collating students' views was a process of reflexive thematic analysis, as explained by Braun and Clarke (2006, 2022). This systematic qualitative analytical technique relies on the researcher being immersed and situated in the learning world of the students, from which their responses spring, and it has proved to be a rewarding way of gaining a rich understanding of the views of students. As a headteacher as well as a researcher, I am keenly interested in the topic, both from an analytical perspective and from my personal commitment to social justice and the moral purpose of educating young people.

The form of this study was, to a large extent, necessarily shaped by the effects on schooling in England of the COVID-19 pandemic. The pilot study had to be conducted via the internet, since it was not possible to visit schools. Patterns of engagement in this study were much thinner than in the main study, which was carried out in person at the school of which I am the headteacher. Students in the main study had also had their education disrupted through two national lockdowns and a further local lockdown. The pandemic caused the cancellation of public examinations for two consecutive years. During this time, opportunities might have been taken by examination boards to rethink and review assessment practices; instead,

examinations resumed with almost no evidence of change apart from some short-term content modifications. In the circumstances, it seems entirely reasonable to wonder about the entrenched nature of high-stakes assessments and whether our society's attachment to closed-book invigilated examinations is the best way of interpreting and presenting a measure of the expertise of our students.

The pilot study revealed that students who had not already engaged deeply with the questions by attempting to answer them showed an understanding of question difficulty that was much less nuanced or well developed than that of the students who answered the questions first. This finding makes intuitive sense – students experience difficulty through their own actual performance rather than in an abstract way – and it might have been expected, given both the author's extensive personal experience as an educator over 25 years and existent research (see, for instance Hacker *et al.*, 2000; Isaacson and Fujita, 2006⁹⁸). Even students who have already answered questions were found not to be very accurate in evaluating the sources and extent of their difficulty. That students are not particularly accurate at predicting or assessing the difficulty of the questions they encounter has important, and ongoing, practical implications for teaching and for students' examination preparation. In particular, it became evident that students tended to make superficial judgements about whether a question 'looks easy', but that these judgements were often based on surface-level question elements, such as the length of the question, how many marks it carried, and how many steps were involved. Given the lack of previous research engaging with students about question-level difficulty, this is new knowledge. In Chapters 5 and 6, links were found between students' self-efficacy and motivation, and their subsequent success in answering questions; these links are explained by Bandura's theory of self-efficacy (1997) and Marzano's and Kendall's New Taxonomy (2007),

⁹⁸ Hacker *et al.* (2000) found that accuracy in predicting test scores varied widely between students, with lowest performing students showing 'gross overconfidence in predictions' (p. 160). Isaacson and Fujita (2006, p. 39) found that 'high achieving students were more accurate in predicting their test results.' Both studies asked undergraduate (and not secondary school) students to predict test difficulty and their own test scores; neither study asked students to predict question difficulty.

where the student's engagement of their self-system was positively correlated with the likelihood that they would be successful in answering the question.

These findings, brought together, may have important implications for improving teaching and learning in the classroom. For example, it might be helpful for teachers to know that students will tend to over-estimate the difficulty of a long question and under-estimate the difficulty of a short question. By using suitably scaffolded examples and questions of a variety of lengths in lessons, teachers could equip students with strategies so that they do not give up on long questions, when they can successfully answer either all or part of the question. Conversely, teachers could train students, by using carefully chosen examples, not to under-estimate the demands of short questions, but to unpack the question demands carefully so that they fully grasp what the examiners are asking them to do.

It might have been expected that there would be a straightforward link between the cognitive load demands imposed by a given examination question and the student's experienced difficulty, such that there was a direct link between the expertise of a student and their ability to answer a question correctly. In Chapters 5 and 6, however, it was discovered that this link was not so straightforward, because of factors that create difficulty for students in ways that are not predicted by examiners. Sources of unexpected difficulty found in this study included distracting information or context in the question; inappropriate focus in the question; a failure of communication, where the question appears to the student to be asking something different from what the examiners intended; and the length of the question, including the amount of text that has to be read. These sources of unexpected difficulty may impose extraneous cognitive load on a student's working memory. Moreover, and as this study has revealed, those students who are "less expert" are particularly susceptible to the unintended consequences of unexpected difficulty; perhaps because their schemas for tackling questions are less well-developed (that is, less well embedded in their long-term memory) and therefore their working memory is already more "clogged" in processing the demands of the question.

This again is a contribution to knowledge. It has important consequences for teaching and learning in the classroom, because it would be helpful for teachers and students to understand that question demands arise from many sources, and not simply from the inherent demands of the topic and the immediately visible response strategies required by the question. If students are able to understand the sources of possible demands in the questions they encounter, and are trained in controlling their metacognitive processes as they deal with unseen challenges, they may be better equipped to meet these demands.

This study revealed, through both its review of existent literature and the analysis of student voice responses, that examination stress may also deplete the working memory capacity of students. The study showed that students who showed a lack of self-efficacy or confidence often found it difficult to engage fully with questions, and were therefore less likely to be successful. It was also found in this study that students' responses linked well with taxonomies of learning, particularly with Marzano's and Kendall's New Taxonomy (2007), and that the Noel Burch Competency Model could provide straightforward and usable insights into the extent to which students were conscious of their own cognitive competence. This is important to note because these taxonomies and models are readily available and they provide insights and frameworks upon which teachers and students can confidently base their learning. Better understanding of these models might help teachers and students to develop metacognitive knowledge and skills.

Central to this study has been the positioning and placement of "student voice". Throughout the later chapters, students' verbatim comments have been quoted; bringing the views of the students to life and causing them to be present in the discussion. Through so doing, it gave them a place in the academic discourse about high-stakes examinations. The depth and range of students' comments presented in this study, and the degree of insight that they represent, indicate that students' views can make a valuable contribution to the evaluation and

improvement of examinations. The systematic inclusion of student voice within professional evaluations of assessments might bring a measure of equity to the examination system, and help ensure that the relationship between students and examiners is more balanced. More widely, this study shows that the voices of students can be welcomed in academic discourse, and can be regarded not as threats but as valuable contributions to a shared endeavour. Through the research methods used in this study, as headteacher-researcher I have gained a practical understanding of effective ways of consulting students, hearing their voices and analysing what they say. These methods have been grounded in existing literature for student voice, developed through creating and testing survey instruments, and further explored through devising and testing open questions within focus groups.

8.4 Recommendations, applications, and contributions to enhancing academic knowledge

Given its various findings and its use of student voice, this study offers a series of recommendations by which the high-stakes assessment system of GCSE examinations might be improved by being made more equitable and more effective.

Recommendation 1: Common Understanding of Terminology. All members of the assessment community – including examiners, teachers and researchers – should understand and use the terms “demand” and “difficulty” precisely and accurately.

This study has shown that examiners appear to be in a powerful position, set apart from other members of the assessment community. Through creating and using a common vocabulary around demand and difficulty, the different members of the assessment community will be able to communicate with one another precisely and without ambiguity. Creating a strong sense of a shared language is recognised by many businesses and corporate trainers as an essential tool for developing the culture of their enterprises. The use of shared vocabulary encompassing shared meaning around demand and difficulty would enable the assessment

community to understand their goals and expectations in creating examinations that have a strong validity argument. It would increase efficiency because of reduced time taken for explanation; and it would improve the quality of assessment culture, because shared language and vocabulary would help build trust and reciprocal understanding between the different members of the assessment community.

Recommendation 2: Student Voice. Examination boards would benefit from the creation of systematic and structured opportunities to listen to the voices and experiences of students in relation to sources of difficulty in GCSE and other public examinations, in order to improve the quality of these high-stakes assessments. Teachers, who also make frequent use of GCSE examination questions in teaching and assessment, would therefore also benefit from understanding students' perspectives about how they experience and comprehend difficulty in the examination questions they encounter, in order to use these questions more intentionally and effectively as tools to develop learning.

As this study has demonstrated, students are able to take a responsible and active part in discussions about how examinations work for them, and they can communicate messages about their experiences of examinations that are unavailable from any other source. Some of these messages indicate that some aspects of the examination system as a whole and features of some individual questions in particular may not work as well as they should. Incorporating a mechanism to listen to and evaluate students' views into the process of question selection and examination evaluation could therefore improve the quality and effective operation of examination questions. This would also better reflect the position of students as the main stakeholders in public examinations – a position which is, at present, conspicuous by the almost complete absence of their voices and experiences. The examination system appears to some to be opaque and “stacked against” students (TES, 2019; BBC, 2020). Through rebalancing the relationship between examiners, researchers and students, therefore, confidence and trust in the system of public examinations could be improved. The practical

realisation of this recommendation would bring the voices of students into the examination system, improving levels of confidence in the system among stakeholders including students, their teachers and parents.

Recommendation 3: Examiner Training. Training offered to senior examiners could be broadened to make sure that all those who write examination questions and who create examination papers and mark schemes have a good grounding in cognitive load theory, so they can ensure their questions focus appropriately on legitimate sources of demand and avoid unintended sources of difficulty. In this way, examiners will be able to manage the intrinsic cognitive load imposed on students and avoid extraneous cognitive load.

Examiners who understood the models of learning presented by cognitive load theory would be better equipped to design questions that focused on the elements they wished particularly to examine. It is legitimate for examiners to set questions that impose high cognitive loads, in order to discriminate effectively between students of different levels of knowledge and expertise, but they should be precisely aware of the cognitive load they are imposing in each question. Examiners who knew, for example, from their understanding of cognitive load theory, that questions requiring several different cognitive elements to be processed simultaneously imposed higher cognitive loads, and that 'mathematical tasks tend to be high in element interactivity' (Sweller *et al.*, 1998, p. 260), would be able to determine more precisely the extent of the cognitive load they wished to impose. They would also be able to identify and avoid possibly unwanted cognitive load effects, such as the split attention effect and the redundancy effect (see Section 3.1.2), so as to optimise intrinsic cognitive load and minimise extraneous cognitive load. Examiners who understood that contextualising questions and supplying additional information frequently confuses students (rather than making their questions accessible) would be able to avoid these unpredictable and distorting effects. Such knowledge and understanding would enable examiners to make stronger links between the

demands they intended to impose and the difficulties experienced by students. This stronger link between intended and experienced difficulty would enable GCSE examination grades more accurately to represent a measure of the knowledge and expertise of students, thus strengthening the validity of inferences made from examination results.

Recommendation 4: Teacher Education. Initial teacher education and continuing professional development for teachers should include a thorough grounding in cognitive load theory, so that teachers better understand the challenges faced by their students, both in lessons and in examinations. Secondly, more emphasis should be placed on the development of effective formative assessment practices as part of standard classroom pedagogy. These changes should enable teachers to adapt their teaching techniques to help their students learn more effectively. Thirdly, teachers and school leaders would benefit from learning about structured ways to harness the power of student voice to help them evaluate the effectiveness of their teaching and learning practices.

Over the past few years, there has been a growing emphasis in the United Kingdom on the desirability of understanding cognitive load theory, both within initial teacher education and in continuing professional development for teachers within the United Kingdom.⁹⁹ Cognitive load theory is now included in initial teacher training and is explicitly referenced within the UK Government's Early Career Framework (DfE 2019). There is currently no equivalent expectation that an understanding of cognitive psychology should underpin the continuous professional development that serving teachers receive, however. Teachers who understand

⁹⁹ Evidence for this comes, for example, from the endorsement by the influential teacher educator Dylan Wiliam (2017): 'I've come to the conclusion Sweller's Cognitive Load Theory is the single most important thing for teachers to know', Twitter, 26.01.2017 <https://twitter.com/dylanwiliam/status/824682504602943489?lang=en-GB> accessed 12.06.2022; the inclusion of a range of articles introducing and applying cognitive load theory to classroom situations in the publications of the Chartered College of Teachers (for example issue 8, Spring 2020, of their publication *Impact* <https://my.chartered.college/impact/issue-8-cognition-and-learning/> accessed 12.06.2022; and the evidential role of cognitive load theory in the Ofsted Education Inspection Framework (Ofsted, 2019).

how students learn, and the factors that can overload their cognitive systems, are better able to design effective sequences of learning, including opportunities for explicit instruction and effective assessment, so that students can learn more effectively. Through a grounding in cognitive load theory, teachers could also help students understand the demands and difficulties implicit in examination questions, so they could help improve their students' abilities to analyse precisely what an examination question is asking.

Interactions with students in this study demonstrate that they found it beneficial and stimulating to be involved in critical reflection on their own assessment and learning. Black and Wiliam (1998) found that, although formative assessment can lead to significant improvements in learning, practice by teachers was weak, with assessments that tended to focus on factual recall without critical reflection. Darling-Hammond *et al.* (2020) re-emphasised this same finding, and showed that improved understanding of cognitive science reinforced the beneficial role of critical reflection in formative assessment. Training teachers to help students critically evaluate the sources and extent of difficulty in the questions they encounter could result in considerably enhanced student learning.

In the course of this study, and while working simultaneously within secondary schools as a headteacher and as a local authority adviser, I have developed my understanding of practical ways in which students' voices can be heard and their insights brought into the evaluation of teaching and learning. As discussed in section 3.4, student voice practices can occupy points on a continuum between tokenistic and transformational, and there are potential pitfalls in giving agency to students to co-develop more effective pedagogy. Nonetheless, there is potential benefit for students and their teachers in developing structured systems to listen carefully to the considered feedback of students. Teachers and school leaders who understand how students learn could then engage students in structured discussions to further develop their pedagogical practices. Structured methods for engaging students in discussions about their

own learning could therefore usefully be incorporated into initial teacher education and continual professional development for teachers.

Recommendation 5: Student Management of Cognitive Load. Students should be taught to understand and apply cognitive load theory principles, in order to manage their own cognitive load in the classroom and in examinations.

Students who understand how they learn best are in a position of considerable control over their own learning. According to the EEF Learning Toolkit, student metacognitive and self-regulation strategies represent ‘very high impact for very low cost based on extensive evidence’ (EEF, 2016), with approaches in mathematics reported as being particularly successful. This metacognitive grasp helps students manoeuvre themselves, within the Noel Burch Competence Model, from the position of unconscious incompetence to that of conscious competence. Students who understand how they learn can see the value of the learning opportunities given to them by effective teaching, which enables them to take better advantage of these learning opportunities; students who understand how they learn can also teach themselves (and one another) more effectively. If students were routinely taught the rudiments of cognitive load theory and other elements of cognitive psychology, they would be better able to manage their own learning processes, optimising the creation of schemas to store knowledge and understanding in their long-term memory and avoiding cognitive overload. Such an understanding of their own cognition and learning processes could also improve students’ levels of self-efficacy. This, in turn, would enable them to perform more effectively within assessments, ultimately improving the link between their levels of understanding and the examination grades they were awarded.

In addition to these five recommendations, this study has also made a unique contribution to the furtherance of existent academic knowledge and understanding through the methods adopted in this study, the position of the researcher, the ways in which it has allowed the authentic voices of students to be heard, and the links it makes between cognitive learning theory and the real-world experience of examinations.

This is practitioner-led research, of which (as was discovered in the review of the literature) there is very little of a robust standard in relation to assessment. As discussed in the methodology chapter, the researcher is a member of the educational and social cultures from which the students are drawn, fully immersed in the world of schools and examinations in which the students operate. This is a strength of the present study: the researcher, by being able to present and interpret students' responses, has become part of the dialogue between theory and practice. This idiographic approach – starting with the students' own views and responses and studying the language they use with a hermeneutic, interpretative methodology – aligns well with this study's constructivist approach, as discussed in the methods chapter.

The research methodology presented here includes a robust qualitative method, where 'data analysis is conceptualised as an art not a science; creativity is central to the process, situated within a framework of rigour' (Braun and Clarke, 2022, p. 8). This inductive approach is unusual, in a field where the significant studies using student data are predominately (and increasingly) quantitative.

This study has made contributions to knowledge and understanding, in the following ways.

First, in terms of **equity and authenticity**: in this study, the voices of students have been central to the research. Students are arguably the major stakeholders in public examinations, but they are not consulted or involved in any way in evaluating or improving the examination system. Students' contributions in this study have been reported in their own words; they have been critically evaluated, and they have been continually related to theoretical frameworks.

From the thematic analysis that has been presented here, it can be seen that students are able to understand and offer views on examinations that hold value and importance.

Secondly, this study contributes to **pedagogical practice**. There have been very few studies previously that have taken students' voices seriously, and almost none that have systematically reported, analysed and evaluated the views of secondary school students in relation to high-stakes assessments. While this study does not "advocate" for students, it does present and critically evaluate their views. In this way it makes a unique contribution to research and to the development of assessment practice and the understanding of teaching and learning.

8.5 Limitations of this study

It is important to acknowledge the limitations of this study, since they may have a bearing on the generalisability of its conclusions and findings. These limitations analyse into the categories of unforeseen circumstances, a lack of previous research, and choices in research design. First, the period of data collection was substantially affected by the consequences of the COVID-19 pandemic on schooling. The practical implications of this have already been discussed. This study chose to look at questions in GCSE mathematics, which may limit its applicability to other subjects and qualifications.

There was an absence of previous studies using student voice in secondary schools to discuss aspects of teaching, learning and assessment. While this provided opportunities for this study to make its own rich contribution to the tapestry of knowledge and understanding, it also meant that there was not an already-established research grounding in the area. Therefore the literature base for this study had to be wide, drawing on adjacent areas.

The technique of reflexive thematic analysis, as developed by Braun and Clarke, may be unfamiliar to some readers. As such, they may interpret some of its features as potential

limitations, rather than seeing them as sources of strength. The qualitative paradigm acknowledges that research takes place in 'an only-partially knowable world, where meaning and interpretation are always situated practices' (Braun and Clarke, 2022, p. 6) and where the researcher is an interpreter of meaning, a subjective storyteller. For those who are more accustomed to a quantitative paradigm, this may make for uneasy reading, because it is far removed from positivist ideas of an objective truth that can be hypothesised, tested and verified. The selection of reflexive thematic analysis as the method for this study, however, was a deliberate and positive choice, and it is a robust and respected research method used increasingly in the social sciences. While opportunities were taken to collaborate in discussing research questions, data collection methods and emergent findings with professional colleagues and supervisors, no claims are made for repeatability or wholesale generalisability. Sample sizes in this study are relatively small: around 100 participants in each part of the study (it may be noted, however, that these sample sizes are relatively large for a qualitative study), and students were drawn from one geographically restricted area in the North East of England. Reflexive thematic analysis is by its nature very time-consuming, which puts a practical limit on the number of participants and the number of times any researcher can conduct such studies. The experiences and insights of these students may not be representative of other populations of students, therefore, and it might be observed that the conclusions from a small set of studies may not be the strongest basis for recommending wholesale changes to national examination practice.

Taking all the above factors into account, it can be seen that it is not possible to repeat this study and expect to obtain the same findings. The view from this thesis, however, is that this is not a weakness, but a desirable feature of situated and immersive qualitative research.

8.6 Avenues of future research

There are a number of avenues of future potential research avenues that build directly upon aspects of this study. First, the two empirical parts of this study were carried out in particular topic areas of mathematics – a “core” subject of the UK National Curriculum – and past paper examination questions were selected from a single examination board. Future research might consider broadening either the subject basis to include other core subjects (English and science), non-core curriculum areas such as geography or history, and to questions from a wider span of examination boards.

Secondly, whilst this study focused on GCSE student cohorts, further research might seek to broaden this by also taking account of the experiences of those of a similar age who are subject to different examination processes; for instance Scottish National 4 or National 5 examinations, or the International GCSE.

Thirdly and finally, to further cement the centrality of student voice in research of this type – of which this thesis may be seen to be a pioneering example – future researchers might consider broadening the methodological approaches adopted to investigate and comparatively evaluate the views of senior examiners and teachers alongside those of students.

8.7 Concluding remarks

In conclusion, this study has introduced fresh perspectives and knowledge, and has contributed to a deeper understanding of the workings of the public examinations system. It has been demonstrated that there is inconsistency in the ways in which the terms “demand” and “difficulty” are used by professional educators and members of the assessment community, and that a thorough understanding of the concepts to which these terms refer is not consistent in research literature. Suggestions are advanced to move this understanding forward.

This study has also shown that secondary school students are able to articulate their understanding of concepts of demand and difficulty in relation to GCSE mathematics examination questions. Their views are coherent and interesting, and they relate well to learning theories such as cognitive load theory and taxonomies of learning such as those of Bloom (as revised by Anderson *et al.*, 2001), and Marzano and Kendall (2007).

This study has also shown how the present relationship between examiners and students appears fundamentally unbalanced within the formal assessment system of public examinations operating in the UK for the award of GCSEs. This imbalance results in a lack of equity, as students are disenfranchised within the high-stakes assessment system of which they are arguably the most important stakeholders. Opportunities to improve the performance of examination questions are, therefore, being missed. Because there is currently no established mechanism for examiners to hear, consider, and act upon the views and experiences of students, examiners cannot know whether the demands they intend to create in their questions become fair and consistent sources of difficulty for students, and they are deprived of opportunities to plan for and avoid unpredictable demands. As a result of these failings within the current model, there are consequential threats to the validity of inferences made from the results of public examinations such as the GCSE, which have far-reaching implications for students, teachers, and schools. Through its research methodology, its engagement with communities of students, and its critical interpretation of students' responses, this study has produced findings and recommendations. These aim to strengthen a shared understanding of the architecture of learning and the vocabulary of assessment within the community of examiners, researchers and teachers, and to mitigate against the unpredictable actions of untested examination questions. Public examinations have an important role in defining and monitoring educational standards. At a time when the role of artificial intelligence appears to threaten the integrity of many formal assessments, it seems likely that closed-book invigilated examinations will continue to be held up as a robust and reliable assessment method. The recommendations of this thesis therefore aim to strengthen

confidence in the examination system. For students and teachers, examinations provide clear goals and purpose at the end of their GCSE course of study, as well as a system of incentives and rewards.

My vision, as headteacher-researcher, reaching the end of this part of a long and fascinating journey, is that implementing the recommendations of this thesis can help to improve the efficacy of the examination system, thereby clarifying and improving its role in the validation and certification of learning. After all, to give the last word to one of the main study students (R51M), *'the question isn't hard as long as you understand what the question's actually asking.'*

Postscript

In August 2023, at the end of this study, I once again found myself with students talking about their examination experiences. My journey had come full circle. But this time I was inside the hall with the students, congratulating them as they collected their A level results. One student had done particularly well, gaining top grades in A level mathematics, further mathematics and physics. He said '*I can't quite believe I've done it*', and he told me he was now heading off to study physics with astrophysics at a prestigious (Russell Group) university, an impressive achievement for any student, and particularly for one who came from a more deprived background. In contrast to his sense of unreality, I could readily believe he had achieved so well. I reminded him that he was one of the students who responded to my lockdown pilot study questionnaire, when he was in Year 10, and that his answers, above those of all other students, had shown not only 'conscious competence' but clear signs of metacognition (see Sections 5.3 and 7.8). I had encouraged him at the time, telling him I saw something special in the depth of his answers. We both smiled at the recollection. This underlines a reason for engaging with students in discussions about their experience and comprehension of demand and difficulty in GCSE questions: critical reflection and metacognition at this stage may indicate and may encourage future high fliers.

This page has been left intentionally blank

References

- Addey, C., Baddox, B, and Zumbo, B.D. (2020). Assembled validity: rethinking Kane's argument-based approach in the context of International Large-Scale Assessments (ILSAs). *Assessment in Education: Principles, Policy and Practice* 27(6), 588-606.
- Agarwal, P. K. (2019). Retrieval practice and Bloom's taxonomy: Do students need fact knowledge before higher order learning? *Journal of Educational Psychology*, 111(2), 189–209.
- Agius, S. J. (2013). Qualitative research: its value and applicability. *The Psychiatrist*, 37(6), 204-206.
- Ahmed, A. and Pollitt, A. (1999). *Curriculum Demands and Question Difficulty*. Paper presented at the International Association for Educational Assessment Conference, Slovenia, May 1999.
- Ahmed, A. and Pollitt, A. (2007). Improving the quality of contextualized questions: an experimental investigation of focus, *Assessment in Education: Principles, Policy and Practice*, 14(2), 201-232.
- Ahmed, A. and Pollitt, A. (2011). Improving marking quality through a taxonomy of mark schemes. *Assessment in Education: Principles, Policy and Practice*, 18(3), 259-278.
- Akulwar-Tajane, I., Raikundlia, H., Gohil, R., and Shinde, S. (2021). Academic stress in physiotherapy students: Are open book examinations the solution in the face of COVID-19 pandemic. *Health Research*, 5(2), 1-28.
- Alderson, J. C. and Wall, D. (1993). Does washback exist? *Applied linguistics*, 14(2), 115-129.
- Alton A and Pearson S (1996). *Statistical Approaches to Inter-Subject Comparability*. Unpublished UCLES research paper.
- Anderson, A. (2015). The critical purchase of genealogy: Critiquing student participation projects. *Discourse: Studies in the Cultural Politics of Education*, 36(1), 42-52.
- Anderson, L.W., Krathwohl, D.R., Airasian, P.W., Cruikshank, K.A., Mayer, R.E., Pintrich, P.R., Raths, J. and Wittrock, M.C. (2001). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Pearson.
- Anthony, D. P., and van Nieuwerburgh, C. J. (2018). A thematic analysis of the experience of educational leaders introducing coaching into schools. *International Journal of Mentoring and Coaching in Education*, 7(4), 343-356.
- Apple, M. W., Au, W., and Gandin, L. A. (2009). Mapping critical education. In Apple, M. W., Au, W., and Gandin, L. A. (Eds.), *The Routledge international handbook of critical education* (3–19). Routledge.
- Ashley, L. D. (2021). Planning Your Research. In: Coe, R., Waring, M., Hedges, L.V. and Ashley, L. D. (2021) *Research Method and Methodologies in Education* (3rd edition). SAGE.

- Atkinson, R. C., and Shiffrin, R. M. (1971). The control of short-term memory. *Scientific American*, 225(2), 82-91.
- Atkinson, R.C., and Shiffrin, R.M. (1968). Human memory: a proposed system and its control processes. In Spence, K.W. and Spence, J.T. (Eds.), *The psychology of learning and motivation: Advances in research and theory*. (Vol. 2, pp. 89–195). New York: Academic Press.
- Baddeley, A. (1992). Working memory. *Science*, 255(5044), 556-559.
- Baddeley, A. D. (2002). Is working memory still working? *European Psychologist*, 7(2), 85–97.
- Baddeley, A.D. and Hitch, G.J. (1974). Working memory. In Bower, G.H. (Ed.), *The psychology of learning and motivation: Advances in research and theory*. Vol. 8. (742–775). New York: Academic Press.
- Baird, J. -A., and Black, P. (2013). Test theories, educational priorities and reliability of public examinations in England. *Research Papers in Education*, 28(1), 5-21.
- Baird, J. -A., Andrich, D., Hopfenbeck, T. N., and Stobart, G. (2017). Assessment and learning: Fields apart? *Assessment in Education: Principles, Policy and Practice*, 24(3), 317-350.
- Baird, J. -A., Chamberlain, S., Meadows, M., Royal-Dawson, L., and Taylor, R. (2009). Students' views of Stretch and Challenge in A-Level Examinations. *Guildford: Assessment and Qualifications Alliance*.
- Bakhtin, M. M. (1981). *The dialogic imagination*. University of Texas Press.
- Ball, S. (2003). The teacher's soul and the terrors of performativity. *Journal of Education Policy*, 18(2), 215–228.
- Ball, S., Maguire, M., and Braun, A. (2012). *How schools do policy: policy enactments in secondary schools*. Routledge.
- Bandura, A. (1986). *Social Foundations of Thought and Action: A Social Cognitive Theory*. Prentice-Hall.
- Bandura, A. (1997). *Self-efficacy: The exercise of control*. Freeman.
- Barbour, R. (2018). *Doing Focus Groups*, 2nd edition. SAGE.
- Barfield, W. (1986). Expert-novice differences for software: Implications for problem-solving and knowledge acquisition. *Behaviour and Information Technology*, 5(1), 15-29.
- Barnard, H.C. (1961). *A History of English Education from 1760*, 2nd edition. University of London Press Ltd.
- Barrance, R. (2019). The fairness of internal assessment in the GCSE: the value of students' accounts, *Assessment in Education: Principles, Policy and Practice*, 26(5), 563-583.
- Barrance, R. and Elwood, J. (2018). National assessment policy reform 14-16 and its consequences for young people: student views and experiences of GCSE reform in

- Northern Ireland and Wales, *Assessment in Education: Principles, Policy and Practice*, 25(3), 252-271.
- Barton, C. (2007): Real Life Maths <https://mrbartonmaths.com/teachers/research/real.html>, accessed 10.04.2023.
- Barton, C. (2014). Comparing Draft GCSE Maths Specifications. Mr Barton Maths Blog. <http://www.mrbartonmaths.com/blog/comparing-draft-gcse-maths-specifications/> accessed 12.06.2022
- Bassey, M. (2001). A Solution to the Problem of Generalisation in Educational Research: Fuzzy prediction, *Oxford Review of Education*, 27(1), 5-22.
- Battista, V., and Cheng, E. (2011). Motion charts: Telling stories with statistics. In *American Statistical Association Joint Statistical Meetings* (Vol. 4473).
- Baumann, J. F., and Duffy, A. M. (2001). Teacher-researcher methodology: Themes, variations, and possibilities. *The Reading Teacher*, 54 (6), 608-615.
- BBC (2019). 'GCSE results: Pass rates and top grades edge upwards', 22 August 2019 <https://www.bbc.co.uk/news/education-49421275> accessed 22 July 2020.
- BBC (2020). 'Scottish school pupils have results upgraded'. BBC News, 11.08.2020. <https://www.bbc.co.uk/news/uk-scotland-53740588> accessed 12.06.2022.
- Beilock, S. L. (2008). Math performance in stressful situations. *Current Directions in Psychological Science*, 17(5), 339–343.
- BERA (2018). British Educational Research Association. Ethical Guidelines for Educational Research, fourth edition. London.
- Berger, R. (2015). Now I see it, now I don't: Researcher's position and reflexivity in qualitative research. *Qualitative research*, 15 (2), 219-234.
- Berger, R. (2018). *Here's What's Wrong with Bloom's Taxonomy: A Deeper Learning Perspective* [blog for Education Week, 14 March, 2018] accessed at <https://www.edweek.org/education/opinion-heres-whats-wrong-with-blooms-taxonomy-a-deeper-learning-perspective/2018/03>.
- Bernstein, B. (2003). *Class, codes and control: The structuring of pedagogic discourse* (3rd edition). New York: Routledge.
- Beveridge, I. (1997). Teaching Your Students to Think Reflectively: the case for reflective journals, *Teaching in Higher Education*, 2(1), 33-43.
- Biggs, J. (1993). What do inventories of students' learning process really measure? A theoretical review and clarification, *British Journal of Educational Psychology*, 83, 3-19.
- Black, P., and William, D. (1998). Assessment and classroom learning. *Assessment in Education: principles, policy and practice*, 5(1), 7-74.

- Black, P. and Wiliam, D. (1998). *Inside the black box: raising standards through classroom assessment*, G L Assessment Limited.
- Black, P., and Wiliam, D. (2018). Classroom assessment and pedagogy. *Assessment in education: Principles, policy and practice*, 25(6), 551-575.
- Black, P., Harrison, C., and Lee, C. (2003). *Assessment for learning: Putting it into practice*. McGraw-Hill Education (UK).
- Blakey, J.M. (2019). Cognitive Load Theory: What Does the Research Say? Blog for *Schools Week*, 25.03.2019. Accessed at <https://schoolsweek.co.uk/cognitive-load-theory-what-does-the-research-say/> 25.02.2022.
- Bloom, B.S., Engelhart, M., Furst, E., Hill, W., and Krathwohl, D. (Eds.) (1956). *Taxonomy of Educational Objectives: The Classification of Educational Goals, Handbook I: Cognitive Domain*. David McKay.
- Boud, D. (2000). Sustainable assessment: rethinking assessment for the learning society. *Studies in continuing education*, 22(2), 151-167.
- Boud, D., and Soler, R. (2016). Sustainable assessment revisited. *Assessment and Evaluation in Higher Education*, 41(3), 400-413.
- Bourke, R. and Loveridge, J. (2018). Using Student Voice to Challenge Understandings of Educational Research, Policy and Practice. In: Bourke, R. and Loveridge, J. (Eds.) *Radical Collegiality through Student Voice*. Springer.
- Bragg, S. (2001). Taking a joke: Learning from the voices we don't want to hear. In *Forum* (Vol. 43:2, 70-73). Symposium Journals.
- Bragg, S. (2007). Student voice and governmentality: The production of enterprising subjects? *Discourse: Studies in the Cultural Politics of Education*, 28(3), 343-358.
- Braun, V. and Clarke, V. (2006). Using thematic analysis in psychology, *Qualitative Research in Psychology*, 3, 77-101.
- Braun, V., and Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative research in sport, exercise and health*, 11(4), 589-597.
- Braun, V., and Clarke, V. (2021a). Can I use TA? Should I use TA? Should I not use TA? Comparing reflexive thematic analysis and other pattern-based qualitative analytic approaches. *Counselling and Psychotherapy Research*, 21(1), 37-47.
- Braun, V., and Clarke, V. (2021b). One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative Research in Psychology*, 18(3), 328-352.
- Braun, V., and Clarke, V. (2022). *Thematic Analysis: A Practical Guide*. SAGE.
- Braun, V., and Clarke, V. (2023). Toward good practice in thematic analysis: Avoiding common problems and be(com)ing a *knowing* researcher. Editorial in *International Journal of Transgender Health*, 24(1), 1-6.

- Broadfoot, P., Murphy, R., and Torrance, H. (2012). The GCSE: promise vs. reality, in Nuttall, D.L. (Ed.) *Changing Educational Assessment* (154-159). Routledge.
- Bruner, J.S. (1967). *On Knowing: Essays for the Left Hand*. New York; Atheneum.
- Burton, L. (1995). The Rights and Wrongs of Teaching Rites of Passage. *British Journal of Religious Education* 17(3), 180-9.
- Cammarota, J. and Fine, M. (2008). *Revolutionizing Education – Youth Participatory Action Research in Motion*, Routledge.
- Carless, D. (2009). Trust, distrust and their impact on assessment reform. *Assessment and Evaluation in Higher Education*, 34(1), 79-89.
- Chamberlain, T., Golden, S. and Bergeron, C. (2011) Children and young people's views of education policy. London: Office of the Children's Commissioner.
- Chandler, P. and Sweller, J. (1991). Cognitive Load Theory and the Format of Instruction, *Cognition and Instruction*, 8(4), 293-332.
- Charteris, J. and Smardon, D. (2019a). The politics of student voice: unravelling the multiple discourses articulated in schools, *Cambridge Journal of Education*, 49(1), 93-110.
- Charteris, J. and Smardon, D. (2019b). Democratic contribution or information for reform? Prevailing and emerging discourses of student voice. *Australian Journal of Teacher Education*, 44(6), 1-18.
- Chen, O., Castro-Alonso, J. C., Paas, F., and Sweller, J. (2018). Extending cognitive load theory to incorporate working memory resource depletion: evidence from the spacing effect. *Educational Psychology Review*, 30(2), 483-501.
- Cheng, L., and Curtis, A. (2004). Washback or backwash: A review of the impact of testing on teaching and learning. *Washback in language testing*, 25-40.
- Christodoulou, D. (2016). *Making Good Progress? The future of Assessment for Learning*. Oxford University Press.
- Coe, R. (2007). Common Examinee Methods' in P. Newton, J. Baird, H. Goldstein, H. Patrick and P. Tymms (Eds) *Techniques for monitoring the comparability of examination standards*. Qualifications and Curriculum Authority.
- Coe, R. (2015). 'What makes great teaching?' Presentation to IB World Regional Conference, Den Haag, Netherlands, 31.10.2015
<https://www.ibo.org/globalassets/events/aem/conferences/2015/robert-coe.pdf>
 accessed 07.05.2022.
- Cohen, L., Manion, L. and Morrison, K. (2011). *Research methods in education*, 7th edition. Routledge.
- Cohen, L., Manion, L. and Morrison, K. (2018). *Research methods in education*, 8th edition. Routledge.

- Cook-Sather, A. (2006). Sound, presence, and power: "Student voice" in educational research and reform. *Curriculum inquiry*, 36(4), 359-390.
- Cook-Sather, A. (2014). Student voice in teacher development. In Meyer, L. (Ed.), *Oxford bibliographies in education*. Oxford University Press.
- Cook-Sather, A. (2018). Tracing the evolution of student voice in educational research. *Radical collegiality through student voice: Educational experience, policy and practice*, 17-38.
- Cook-Sather, A. (2020). "Respecting voices: How the co-creation of teaching and learning can support academic staff, underrepresented students, and equitable practices." *Higher Education*, 79(5), 885-901.
- Cooper, G., and Sweller, J. (1987). Effects of schema acquisition and rule automation on mathematical problem-solving transfer. *Journal of educational psychology*, 79(4), 347-362.
- Crisp, V. and Grayson, R. (2013). Modelling question difficulty in an A level physics examination, *Research Papers in Education*, 28(3), 346-372.
- Crisp, V. and Macinska, S. (2020). Accessibility in GCSE Science exams – Students' perspectives. *Research Matters: A Cambridge Assessment publication*, 29, 2-10.
- Crisp, V. and Sweiry, E. (2005). 'Can a picture ruin a thousand words? The effects of visual resources in examination questions', *Research Matters: A Cambridge Assessment publication*, 1, 11-15.
- Cronbach, L. J. (1971). Test validation, in Thorndike, R.L. (Ed.) *Educational measurement* (2nd edition, 443-507). Washington, DC: American Council on Education.
- Cronbach, L. J., and Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological bulletin*, 52(4), 281.
- Crotty, M. (1998). *The Foundations of Social Research: Meaning and Perspective in the Research Process*. Sage Publications.
- Crowe, A., Dirks, C., and Wenderoth, M. P. (2008). Biology in bloom: implementing Bloom's taxonomy to enhance student learning in biology. *CBE—Life Sciences Education*, 7(4), 368-381.
- Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (621–694). American Council on Education.
- Darling-Hammond, L., Flook, L., Cook-Harvey, C., Barron, B., and Osher, D. (2020). Implications for educational practice of the science of learning and development. *Applied developmental science*, 24(2), 97-140.
- De Carvalho, E., and Skipper, Y. (2019). "We're not just sat at home in our pyjamas!": a thematic analysis of the social lives of home educated adolescents in the UK. *European Journal of Psychology of Education*, 34(3), 501-516.

- De Groot, A. D. (1966). Perception and memory versus thought: Some old ideas and recent findings. In Kleinmuntz, B. (ed.), *Problem Solving*, Wiley, New York.
- De Jong, T. (2010). Cognitive load theory, educational research, and instructional design: Some food for thought. *Instructional science*, 38(2), 105-134.
- Dearing, R. (1996). *Review of Qualifications for 16-19 year olds*. Schools Curriculum and Assessment Authority.
- DeCoster, J. (2006). Testing Group Differences using T-tests, ANOVA, and Nonparametric Measures. Retrieved 01.06.2022 from <http://www.stathelp.com/notes.html>
- Dempster, E. R., and Kirby, N. F. (2018). Inter-rater agreement in assigning cognitive demand to Life Sciences examination questions. *Perspectives in Education*, 36(1), 94-110.
- DfE (2014a): Listening to and involving children and young people.
https://dera.ioe.ac.uk/19522/1/Listening_to_and_involving_chidren_and_young_people.pdf accessed 02.06.2022.
- DfE (2014b). National Curriculum in England: Mathematics Programmes of Study.
<https://www.gov.uk/government/publications/national-curriculum-in-england-mathematics-programmes-of-study/national-curriculum-in-england-mathematics-programmes-of-study> accessed 12.06.2022
- DfE (2019). Early Career Framework.
https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/978358/Early-Career_Framework_April_2021.pdf updated April 2021. Accessed 12.06.2022.
- DfE (2020). Actions for Schools during the Coronavirus Outbreak. Retrieved 01.06.2022 from <https://www.gov.uk/government/publications/actions-for-schools-during-the-coronavirus-outbreak> Guidance withdrawn February 2022.
- DfE (2023). The Education Hub: SATs English Reading Test – were they Year 6 tests more difficult in 2023 than previous years?
<https://educationhub.blog.gov.uk/2023/05/18/sats-english-reading-test-were-the-year-6-tests-more-difficult-in-2023-than-previous-years/> accessed 14.07.2023
- Dhillon, D. (2003). *Predictive models of question difficulty – A critical review of the literature*. Manchester: AQA Centre for Education Research and Policy.
- Dhillon, D. and Richardson, M. (2003). Another difficult question? An investigation of problem solving and question difficulty issues concerning gifted and talented students. *Paper presented at BERA, Herriot-Watt University, Edinburgh, September 2003*.
- Didau, D. (2019). *Making Kids Cleverer: A Manifesto for Closing the Advantage Gap*. Crown House Publishing Limited.
- Dockerill, B. (2018). 'Forgotten Voices': The Debating Societies of Durham and Liverpool, 1900-1939. In: Burkett, J. (ed.) *Students in Twentieth-Century Britain and Ireland*. Palgrave Macmillan.

- Dolin, J., Black, P., Harlen, W., and Tiberghien, A. (2018). Exploring relations between formative and summative assessment. In Dolin, J. and Evans, R. (Eds) *Transforming assessment*, 53-80. Springer.
- Dunleavy, A., and Sorte, R. (2022). A thematic analysis of the family experience of British mainstream school SEND inclusion: can their voices inform best practice? *Journal of Research in Special Educational Needs*, 22(4), 332-342.
- Edwards, J., and Dall'Alba, G. (1981). Development of a scale of cognitive demand for analysis of printed secondary science materials. *Research in Science Education*, 11, 158–170.
- EEF (2016). The Sutton Trust-Education Endowment Foundation Teaching and Learning Toolkit. <https://educationendowmentfoundation.org.uk/education-evidence/teaching-learning-toolkit/metacognition-and-self-regulation> accessed 12.06.2022.
- EEF (2018). Metacognition and Self-regulated Learning. <https://educationendowmentfoundation.org.uk/education-evidence/guidance-reports/metacognition> accessed 14.07.2023
- Ellerton, P. (2022). On critical thinking and content knowledge: A critique of the assumptions of cognitive load theory. *Thinking Skills and Creativity*, 43, 100975.
- Elliott, R., Fischer, C. T., and Rennie, D. L. (1999). Evolving guidelines for publication of qualitative research studies in psychology and related fields. *British Journal of Clinical Psychology*, 38 (3), 215-229.
- Elliott, T. (1994): Making strange what had appeared familiar. *The Monist*, 77(4), 424-433.
- Elwood, J. (2012). Qualifications, examinations and assessment: Views and perspectives of students in the 14–19 phase on policy and practice. *Cambridge Journal of Education*, 42(4), 497–512.
- Entwistle, N. (1991). Approaches to learning and perceptions of the learning environment: Introduction to the special issue. *Higher Education*, 22, 201-204.
- Erickson, S., and Heit, E. (2015). Metacognition and confidence: comparing math to other academic subjects. *Frontiers in Psychology*, 6, 742.
- Esbensen, F. A., Melde, C., Taylor, T. J., and Peterson, D. (2008). Active parental consent in school-based research: how much is enough and how do we get it? *Evaluation review*, 32(4), 335-362.
- Fautley, M. (2015). Music Education, Assessment and Social Justice: Resisting Hegemony Through Formative Assessment, in Benedict, C., Schmidt, P., Spruce, G. and Woodford, P. (Eds) *The Oxford Handbook of Social Justice in Music Education*, Oxford University Press.
- Fautley, M. and Savage, J. (2008). *Assessment for Learning and Teaching in Secondary Schools*. Learning Matters Ltd.
- Field, A. (2009). *Discovering statistics using SPSS*, 3rd edition. Sage Publications.

- Fielding, M. (2001). Beyond the rhetoric of student voice: New departures or new constraints in the transformation of 21st century schooling? In *Forum for promoting 3-19 comprehensive education* 43(2), 100-109.
- Fisher-Hoch, H., and Hughes, S. (1996). What makes mathematics exam questions difficult? *British Educational Research Association, University of Lancaster, England*, 66.
- Florian, L., and Beaton, M. (2018). Inclusive pedagogy in action: getting it right for every child. *International journal of inclusive education*, 22(8), 870-884.
- Forehand, M. (2010). Bloom's taxonomy. *Emerging perspectives on learning, teaching, and technology*, 41(4), 47-56.
- Fosnot, C.T. (2013). 'Preface'. In Fosnot, C.T. (ed.), *Constructivism: theory, perspectives and practice*, (2nd edition, 1-3). Teachers College Press.
- Francis, B., Skelton, C., and Read, B. (2010). The simultaneous production of educational achievement and popularity: How do some pupils accomplish it? *British Educational Research Journal*, 36(2), 317-340.
- Fuller, E., Rabin, J. M., and Harel, G. (2011). Intellectual need and problem-free activity in the mathematics classroom. *Jornal Internacional de Estudos em Educação Matemática*, 4(1).
- Furst, E. (1994). 'Bloom's Taxonomy: Philosophical and Educational Issues.' In Anderson, L. and Sosniak, L. (Eds.) *Bloom's Taxonomy: A Forty-Year Retrospective*, 28-40. Chicago: The National Society for the Study of Education.
- Geary, D. C. (1995). Reflections of evolution and culture in children's cognition: Implications for mathematical development and instruction. *American psychologist*, 50(1), 24.
- Geary, D. C., and Berch, D. B. (2016). Evolution and children's cognitive and academic development. In *Evolutionary perspectives on child development and education*, 217-249. Springer.
- Gibbs, A. (2021). Focus Groups and Group Interviews. In Coe, R., Waring, M., Hedges, L.V., and Ashley, L.D. (Eds.) (2021). *Research Methods and Methodologies in Education*, 3rd edition. SAGE.
- Giddens, A. (1979). *Central Problems in Social Theory*. London: Macmillan.
- Goldstein, H. and Cresswell, M. (1996). The comparability of different subjects in public examinations: a theoretical and practical critique, *Oxford Review of Education*, 22(4), 435-442.
- Goodenow, C. and Grady, K.E. (1993). The relationship of school belonging and friends' values to academic motivation among urban adolescent students. *The Journal of Experimental Education*, 62(1), 60-71.
- Gray, L. and Baird J. (2020). Systemic influences on standard setting in national examinations, *Assessment in Education: Principles, Policy and Practice*, 27(2), 137-14.

- Greatorex, J., Ireland, J. and Coleman, V. (2019). Two taxonomies are better than one: towards a method of analysing a variety of domains and types of thinking in an assessment. *Educational futures*, 10(1), 3-30.
- Greatorex, J., Shaw, S., Hodson, P., Ireland, J., and Werno, M. (2013). Using scales of cognitive demand in a validation study of Cambridge International A and AS Level Economics. *Research Matters: A Cambridge Assessment Publication*, 15, 29-37.
- Groundwater-Smith, S. and Mockler, N. (2016). From data source to co-researchers? Tracing the shift from 'student voice' to student-teacher partnerships in Educational Action Research. *Educational Action Research*, 24(2), 159-176.
- Guardian (2012). 45,000 Resit GCSE English Exams
<https://www.theguardian.com/education/2012/oct/11/45000-resit-gcse-english-exams> accessed 14.07.2023
- Guardian (2023). Headteachers express concern over SATs amid claims a paper left pupils in tears <https://www.theguardian.com/education/2023/may/11/headteachers-express-concern-over-sats-amid-claims-a-paper-left-pupils-in-tears> accessed 04.07.2023
- Guardian, (2011). 'Criticism over plans to exclude religious studies from Ebacc,' 25 June 2011
<https://www.theguardian.com/education/2011/jun/25/religious-studies-ebacc-exclusion> accessed 19.01.2022.
- Guba, E. and Lincoln, Y. S. (1998). 'Competing Paradigms in Qualitative Research.' In Denzin, N. K. (ed.) *The Landscape of Qualitative Research: Theories and issues*, 195–220. Sage Publications.
- Gunter, H., and Thomson, P. (2007). Learning about student voice. *Support for learning*, 22(4), 181-188.
- Hacker, D. J., Bol, L., Horgan, D. D., and Rakow, E. A. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology*, 92(1), 160-170.
- Halcomb, E., Gholizadeh, L., DiGiacomo, M., Phillips, J., and Davidson, P. (2007). Literature review: considerations in undertaking focus group research with culturally and linguistically diverse groups, *Journal of Clinical Nursing*, 16(6), 1000-11.
- Hall, V. (2017). A tale of two narratives: student voice – what lies before us? *Oxford Review of Education*, 43(2), 180-193.
- Hambleton, R. K., and Swaminathan, H. (1985). Item Banking. In: Hambleton, R. K., and Swaminathan, H. (eds.) *Item Response Theory*. Springer
- Hamer, J., Murphy, R., Mitchell, T., Grant, A., and Smith, J. (2013). *English Baccalaureate Certificate (EBC) proposals: Examining with and without tiers*. Pearson.
- Hammersley, M. (1992). *What's wrong with ethnography?* London: Routledge.
- Hanushek, E. A., and Wößmann, L. (2008). *Education quality and economic growth*. World Bank.

- Hariton, E., and Locascio, J. J. (2018). Randomised controlled trials—the gold standard for effectiveness research. *BJOG: an international journal of obstetrics and gynaecology*, 125(13), 1716.
- He, Q., Opposs, D., Glanville, M., and Lampreia-Carvalho, F. (2015). Assessing pupils at the age of 16 in England—approaches for effective examinations. *Curriculum Journal*, 26(1), 70-90.
- Heller-Sahlgren, G. (Ed.) (2014). *Testing worth teaching to – incentivising quality in qualifications and accountability*. The Centre for Market Reform of Education.
- Herrington, J., and Standen, P. (2000). Moving from an instructivist to a constructivist multimedia learning environment, *Journal of Educational Multimedia and Hypermedia*, 9(3), 195–205.
- Hillman, N. (2014). A guide to the removal of the student number controls. *HEPI Report 69*: Higher Education Policy Institute.
- Hohnen, B. and Murphy, T. (2016). The optimum context for learning: drawing on neuroscience to inform best practice in the classroom *Educational and Child Psychology*, 33(1), 75-90.
- Hopfenbeck, T.N. (2019). Assessment reforms and grading, *Assessment in Education: Principles, Policy and Practice*, 26(3), 255-25.
- Hopfenbeck, T.N. (2020). Rethinking validity in educational assessment, *Assessment in Education: Principles, Policy and Practice*, 27(6), 585-587.
- Hopkins, E. A. (2008). Classroom conditions to secure enjoyment and achievement: the pupils' voice. Listening to the voice of Every child matters. *Education 3–13*, 36(4), 393-401.
- Howard, T. C. (2001). Telling their side of the story: African-American students' perceptions of culturally relevant teaching. *The Urban Review*, 33(2), 131-149.
- Hughes, S., Pollitt, A., and Ahmed, A. (1998). The development of a tool for gauging the demands of GCSE and A Level exam questions. *Paper presented at BERA, Queen's University Belfast*.
- Iannone, P. and Simpson, A. (2015). Students' preferences in undergraduate mathematics assessment. *Studies in Higher Education*, 40(6), 1046-1067.
- Impara, J. C. and Plake, B. S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement*, 35(1), 69-81.
- Irvine, J. (2017). A Comparison of Revised Bloom and Marzano's New Taxonomy of Learning. *Research in Higher Education Journal*, 33.
- Isaacson, R., and Fujita, F. (2006). Metacognitive knowledge monitoring and self-regulated learning. *Journal of the Scholarship of Teaching and Learning*, 6(1), 39-55.

- ISM (2019). *Music Education: State of the Nation*, All Party Parliamentary Group on Music, Incorporated Society of Musicians, University of Sussex
<https://www.ism.org/images/images/FINAL-State-of-the-Nation-Music-Education-for-email-or-web-2.pdf>, accessed 19.01.2022.
- Jackson, C. (2010). Fear in Education. *Educational Review*, 62(1), 39-52.
- Jackson, C. (2017). Fear of failure. In *Understanding Learning and Motivation in Youth* (30-39). Routledge.
- Jackson, M., and Lismore-Burns, R. (2012). 'Exploration of questions from GCSE Maths A: Identifying the types of question that candidates find the most and least demanding, within a topic', *AQA Centre for Education research and Policy*
<https://research.aqa.org.uk/research-library/exploration-questions-gcse-maths-identifying-types-question-candidates-find-most-and-least-demanding-within-topic>
 accessed 05.07.2020.
- Jeffries, R., Turner, A. P., and Poison, P. G. P., and Atwood, M. (1981). The processes involved in designing software. *Cognitive skills and their acquisition*, 255-284.
- Johnson, M. and Mehta, S. (2011). Evaluating the CRAS Framework: Development and Recommendations, *Research Matters: A Cambridge Assessment Publication*, 12, 27-33.
- Jones, I., Swan, M., and Pollitt, A. (2015). Assessing mathematical problem solving using comparative judgement. *International Journal of Science and Mathematics Education*, 13(1), 151-177.
- Jones, M. G., and Ennes, M. (2018). *High-stakes testing*. Oxford University Press.
- Kagan, S. (2005). Rethinking Thinking – Does Bloom’s Taxonomy Align with Brain Science? San Clemente, CA: Kagan Publishing. Kagan Online Magazine, Autumn 2005. Accessed at www.kaganonline.com, 05.07.2020.
- Kalyuga, S. (2011). Cognitive Load Theory: How Many Types of Load Does It Really Need? *Educational Psychology Review*, 23, 1–19.
- Kane, M. (2012). All validity is construct validity. Or is it? *Measurement: Interdisciplinary Research and Perspective*, 10(1-2), 66-70.
- Kane, M. (2016), Explicating validity. *Assessment in Education: Principles, Policy and Practice*, 23(2), 198-211.
- Keddie, A. (2015). Student voice and teacher accountability: Possibilities and problematics. *Pedagogy, Culture and Society*, 23(2), 225-244.
- Kehoe, I. (2015). The cost of performance? Students' learning about acting as change agents in their schools. *Discourse: Studies in the Cultural Politics of Education*, 36(1), 106-119.
- Kellaghan, T., and Greaney, V. (2019). *Public examinations examined*. World Bank Publications.
- Kelley, T.L. (1924). Note on the Reliability of a Test: A Reply to Dr. Crum's Criticism. *Journal of Educational Psychology*, 15(4), 193–204.

- Kelly, A.L. (2019). *The High Stakes of Testing*, Brill, Leiden, Netherlands.
- Kidder, L.H., and Fine, M. (1987). Qualitative and quantitative methods: When stories converge. In Mark, M.M. and Shotland, L. (Eds.), *New directions for program evaluation* (57-75). Jossey-Bass.
- Kim, L. E., and Asbury, K. (2020). 'Like a rug had been pulled from under you': The impact of COVID-19 on teachers in England during the first six weeks of the UK lockdown. *British Journal of Educational Psychology*, 90(4), 1062-1083.
- Kim, L. E., Leary, R., and Asbury, K. (2021). Teachers' narratives during COVID-19 partial school reopenings: An exploratory study. *Educational Research*, 63(2), 244-260.
- Kirschner, P. A., and Hendrick, C. (2020). *How Learning Happens: Seminal Works in Educational Psychology and What They Mean in Practice*. Routledge.
- Koretz, D. (2008). *Measuring Up: What Educational Testing Really Tells Us*, Harvard University Press.
- Krathwohl, D. R. (2002). A Revision of Bloom's Taxonomy: An Overview. *Theory Into Practice*, 41(4), 212-218.
- Lakoff, G. (1987). *Women, Fire, and Dangerous Things*. Chicago: University of Chicago Press.
- Lakoff, G. and Turner, M. (1989). *More than Cool Reason: A Field Guide to Poetic Metaphor*. Chicago: Chicago University Press.
- Lau, N. T. T., Hawes, Z., Tremblay, P., and Ansari, D. (2022). Disentangling the individual and contextual effects of math anxiety: A global perspective. *Proceedings of the National Academy of Sciences of the United States of America*, 119(7).
- Lee, C. and Johnston-Wilder, S. (2018). Getting into and staying in the Growth Zone. NRICH. Open University, accessed at <http://oro.open.ac.uk/57155/1/Getting%20into%20and%20staying%20in%20the%20Growth%20Zone.pdf> 29.12.2021.
- Leenknecht, M. J., and Prins, F. J. (2018). Formative peer assessment in primary school: the effects of involving pupils in setting assessment criteria on their appraisal and feedback style. *European Journal of Psychology of Education*, 33, 101-116.
- Lemov, D. (2010). *Teach Like A Champion: 49 Techniques that Put Students on the Path to College* (1st edition), Jossey Bass.
- Lemov, D. (2017). Bloom's Taxonomy – That Pyramid is a Problem. Uncommon Schools Field Notes, 04.03.2017 (*Teach Like A Champion* blog), <https://teachlikeachampion.com/blog/blooms-taxonomy-pyramid-problem/> accessed 27.03.2022.
- Likourezes (2021). An Introduction to Cognitive Load Theory. Blog for The Education Hub. <https://theeducationhub.org.nz/an-introduction-to-cognitive-load-theory/> accessed 23.03.2021.

- Lincoln, Y.S. and Guba, E. (1985). *Naturalistic enquiry*. Sage Publications.
- Little, C. (2008). The role of context in linear equation questions: utility or futility? *Proceedings of the British Society for Research into Learning Mathematics* 28(2).
- Little, C. and Jones, K. (2010). The effect of using real world contexts in post-16 mathematics questions. In Joubert, M. and Andrews, P. (Eds) *Proceedings of the British Congress for Mathematics Education*, 137-144.
- Lodge, C. (2005). From hearing voices to engaging in dialogue: Problematising student participation in school improvement. *Journal of Educational Change*, 6(2), 125–146.
- Lovell, O. (2020). *Sweller's Cognitive Load Theory in Action*. John Catt Educational Ltd.
- Lundy, L. (2007). Voice is not Enough: Conceptualizing Article 12 of the United Nations Convention on the Rights of the Child. *British Educational Research Journal* 33(6), 927–942.
- Lynass, R., Pykhtina, O., and Cooper, M. (2012). A thematic analysis of young people's experience of counselling in five secondary schools in the UK. *Counselling and Psychotherapy Research*, 12(1), 53-62.
- MacBeath, J. (2006). Finding a voice, finding self. *Educational Review* 58(2), 195-207.
- Marcus, N., Cooper, M., and Sweller, J. (1996). Understanding instructions. *Journal of Educational Psychology*, 88, 49–63.
- Martin-Denham, S. (2021). Defining, identifying, and recognising underlying causes of social, emotional and mental health difficulties: thematic analysis of interviews with headteachers in England. *Emotional and behavioural difficulties*, 26(2), 187-205.
- Marzano, R.J. (2001). *Designing a New Taxonomy of Educational Objectives*, Sage Publications.
- Marzano, R.J. and Kendall, J.S. (2007). *The New Taxonomy of Educational Objectives* (2nd edition). Corwin Press (Sage Publications).
- Matthews, K. E., and Dollinger, M. (2023). Student voice in higher education: the importance of distinguishing student representation and student partnership. *Higher Education*, 85(3), 555-570.
- Mavilidi, M. F., and Zhong, L. (2019). Exploring the development and research focus of cognitive load theory, as described by its founders: Interviewing John Sweller, Fred Paas, and Jeroen van Merriënboer. *Educational Psychology Review*, 31(2), 499-508.
- Maxwell, J.A. (1992). Understanding and validity in qualitative research. *Harvard Educational Review*, 62(3), 279-300.
- Maybin, J. (2013). Towards a sociocultural understanding of children's voice. *Language and Education*, 27(5), 383-397.
- McArthur, J. and Huxham, M. (2011). *Sharing Control: A partnership approach to curriculum design and delivery – ESCalate*. Higher Education Academy Education Subject Centre.

- McCormick, R. and James, M. (1983). *Curriculum Evaluation in Schools*. Croom Helm
- Mccrea, P. (2020). *Motivated Teaching*. Amazon
- McEwen, N. (1995). Educational accountability in Alberta. *Canadian Journal of Education*, 20, 27-44.
- McGeechan, G. J., Richardson, C., Wilson, L., Allan, K., and Newbury-Birch, D. (2019). Qualitative exploration of a targeted school-based mindfulness course in England. *Child and Adolescent Mental Health*, 24(2), 154-160.
- McIntyre, D., Pedder, D., and Rudduck, J. (2005). Pupil voice: Comfortable and uncomfortable learnings for teachers. *Research Papers in Education*, 20(2), 149–168.
- Meeran, S., and Davids, M. N. (2022). Covid-19 catalysing assessment transformation: a case of the online open book examination. *South African Journal of Higher Education*, 36(3), 109-122.
- Mendes, A. B., and Hammett, D. (2023). The new tyranny of student participation? Student voice and the paradox of strategic-active student-citizens. *Teaching in Higher Education*, 28(1), 164-179.
- Messick, S. (1989). Validity. In Linn, R.L. (ed.) *Educational measurement*, (3rd edition, 13-103). New York: American Council on Education/Macmillan.
- Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2), 81-97.
- Mitra, D. (2001). Opening the Floodgates: giving students a voice in school reform, *FORUM*, 43(2), 91-94.
- Mitra, D. (2006). Increasing student voice and moving toward youth leadership. *The prevention researcher*, 13(1), 7-10.
- Mitra, D. (2018). Student voice in secondary schools: the possibility for deeper change, *Journal of Educational Administration*, 56(5), 579-487.
- Mitra, D., Serriere, S., and Stoicovy, D. (2012). The Role of Leaders in Enabling Students Voice. *Management in Education* 26(3), 104–112.
- Moreno, R. (2006). When worked examples don't work: Is cognitive load theory at an impasse? *Learning and Instruction*, 16(2), 170-181.
- Murphy, R. (2022). How children make sense of their permanent exclusion: a thematic analysis from semi-structured interviews. *Emotional and Behavioural Difficulties*, 27(1), 43-57.
- Nadar, S. (2014). 'Stories Are Data with Soul'—Lessons from Black Feminist Epistemology. *Agenda* 28(1), 18–28. <https://doi.org/10.1080/10130950.2014.871838>
- Naidoo, R., and Jamieson, I. (2005). Knowledge in the marketplace: The global commodification of teaching and learning in higher education. In Jones, E. and Brown, S. (Eds) *Internationalizing Higher Education*, Springer, 37-51.

- NASUWT advice: Student Voice. <https://www.nasuwt.org.uk/advice/in-the-classroom/children-and-young-people/student-voice.html>, accessed 15.6.2020.
- Newton PE (1997). Measuring Comparability of Standards between Subjects: why our statistical techniques do not make the grade. *British Educational Research Journal*, 23(4), 433-449.
- Newton, P.E. (2012). Clarifying the Consensus Definition of Validity. *Measurement: Interdisciplinary Research and Perspective*, 10(1-2), 1-29.
- Nikitas, A., Wang, J. Y., and Knamiller, C. (2019). Exploring parental perceptions about school travel and walking school buses: A thematic analysis approach. *Transportation research part A: policy and practice*, 124, 468-487.
- Nisbet, I. and Shaw, S. (2020). *Is Assessment Fair?* SAGE publications.
- Novakovic, N., and Greatorex, J. (2011). Comparing the demand of syllabus content in the context of vocational qualifications: literature, theory and method. *Research Matters: A Cambridge Assessment Publication*, 11, 25-32.
- Nowell, L. S., Norris, J. M., White, D. E., and Moules, N. J. (2017). Thematic analysis: Striving to meet the trustworthiness criteria. *International journal of qualitative methods*, 16(1), 1-13.
- Oates, T. (2020). Foreword, *Research Matters: A Cambridge Assessment publication*, 29, 1.
- OCR (2018a). GCSE (9-1) Mathematics Paper 1 (Foundation Tier) Thursday 24 May 2018 – Morning.
- OCR (2018b). GCSE (9-1) Mathematics Paper 4 (Higher Tier) Thursday 24 May 2018 – Morning.
- OCR (2018c). GCSE (9-1) Examiners' report Mathematics J560/01 Summer 2018 series Version 1.
- OCR (2018d). GCSE (9-1) Examiners' report Mathematics J560/04 Summer 2018 series Version 1.
- OCR (2020a). GCSE (9-1) Mathematics Paper 1 (Foundation Tier) Tuesday 03 November 2020 – Morning.
- OCR (2020b). GCSE (9-1) Examiners' report Mathematics J560/01 Autumn 2020 series.
- OCR (2021). GCSE *Mathematics GCSE Specification J560*, Version 1.4 <https://ocr.org.uk/Images/168982-specification-gcse-mathematics.pdf> accessed 22.09.2021.
- Ofsted (2019). Education inspection framework Overview of research. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/963625/Research_for{EIF_framework_updated_references_22_Feb_2021.pdf accessed 12.06.2022.

- Ofsted (2022). School Inspection Handbook.
<https://www.gov.uk/government/publications/school-inspection-handbook-eif/school-inspection-handbook> accessed 12.07.2023
- Ogden, N (2015). New mathematics tiering and content, Blog for OCR.
<https://www.ocr.org.uk/blog/new-gcse-9-1-mathematics-tiering-and-content-shifts>,
 accessed 12.06.2022
- Onwuegbuzie, A.J. and Leech, N.L. (2006). Validity and qualitative research: an oxymoron?
Quality and Quantity, 41(2), 233-49.
- Orr, S. (2010). Collaborating or fight for the marks? Students' experiences of group work
 assessment in the creative arts. *Assessment and Evaluation in Higher Education*, 35(3),
 301-313.
- Oswald, D., Sherratt, F., and Smith, S. (2014). Handling the Hawthorne effect: The challenges
 surrounding a participant observer. *Review of social studies*, 1(1), 53-73.
- Paas, F., Renkl, A. and Sweller, J. (2003). Cognitive Load Theory and Instructional Design:
 Recent Developments. *Educational Psychologist*, 38(1), 1-4.
- Pajares, F., and Kranzler, J. (1995). Self-efficacy beliefs and general mental ability in
 mathematical problem-solving. *Contemporary educational psychology*, 20(4), 426-443.
- Parker, P.D., Van Zanden, B., and Parker, R.B. (2018). Girls get smart, boys get smug: Historical
 changes in gender differences in math, literacy, and academic social comparison and
 achievement, *Learning and Instruction*, 54, 125-137.
- Pearce, T. C., and Wood, B. E. (2019). Education for transformation: An evaluative framework
 to guide student voice work in schools. *Critical Studies in Education*, 60(1), 113-130.
- Pearson, P. D., and Gallagher, M. C. (1983). The instruction of reading
 comprehension. *Contemporary educational psychology*, 8(3), 317-344.
- Perry, T., Lea, R., Jørgensen, C. R., Cordingley, P., Shapiro, K., and Youdell, D. (2021). Cognitive
 Science in the Classroom. Education Endowment Foundation (EEF).
- Piaget, J. (1953). *The Origin of Intelligence in the Child*. Routledge and Keegan Paul.
- Pickett, K., Rietdijk, W., Byrne, J., Shepherd, J., Roderick, P., and Grace, M. (2017). Teaching
 health education: A thematic analysis of early career teachers' experiences following
 pre-service health training. *Health Education*, 117(3), 323-340.
- Pile, S. (2010). Intimate distance: the unconscious dimensions of the rapport between
 researcher and researched. *The Professional Geographer*, 62(4), 483-495.
- Pollitt, A. (2012). The method of adaptive comparative judgement. *Assessment in Education:
 principles, policy and practice*, 19(3), 281-300.
- Pollitt, A., Ahmed, A. and Crisp, V. (2007). 'The demands of examination syllabuses and
 question papers' in Newton, P., Baird, J.A., Goldstein, H., Patrick, H. and Tymms, P.
 (eds.) (2007) *Techniques for monitoring the comparability of examination standards*,

QCA accessed at

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/487056/2007-comparability-exam-standards-g-chapter5.pdf.

- Pollitt, A., Entwistle, N., Hutchinson, C., and de Luca, C. (1985). *What Makes Exam Questions Difficult?* Edinburgh: Scottish Academic Press.
- Posner, D. (2004). What's Wrong with Teaching to the Test? *Phi Delta Kappan*, 85(10), 749-751.
- Powell, M. A., Fitzgerald, R. M., Taylor, N., and Graham, A. (2012). International literature review: Ethical issues in undertaking research with children and young people. *Childwatch International Research Network*.
- Puttick, S. (2015). Chief examiners as Prophet and Priest: relations between examination boards and school subjects, and possible implications for knowledge. *The Curriculum Journal*, 26(3), 468-487.
- Putwain, D. W., and Aveyard, B. (2018). Is perceived control a critical factor in understanding the negative relationship between cognitive test anxiety and examination performance? *School Psychology Quarterly*, 33(1), 65-74.
- Putwain, D. W., and Symes, W. (2018). Does increased effort compensate for performance debilitating test anxiety? *School Psychology Quarterly*, 33(3), 482-491.
- Putwain, D. W., Connors, L., Woods, K. and Nicholson, L. J. (2012). Stress and anxiety surrounding forthcoming standard assessment tests in English schoolchildren. *Pastoral Care in Education*, 30(4), 289–302.
- QCA (2008). *Review of question paper setting and senior examiner training for GCSE and A level*.
https://dera.ioe.ac.uk/9401/1/QCA_08_3581_Question_paper_setting_and_examiner_training_report.pdf accessed 22.09.2021.
- Ramazanoglu, C., and Holland, J. (2002). Women's sexuality and men's appropriation of desire. In *Up against Foucault* (249-274). Routledge.
- Ramsden, P. (1992). *Learning to teach in higher education*. Routledge.
- Rapke, T. (2016). A process of students and their instructor developing a final closed-book mathematics exam. *Research in Mathematics Education*, 18(1), 27-42.
- Rapplee, J., and Komatsu, H. (2018). Stereotypes as Anglo-American exam ritual? Comparisons of students' exam anxiety in East Asia, America, Australia, and the United Kingdom. *Oxford Review of Education*, 44(6), 730-754.
- Reilly, D. (2012) The Four Stages of Competence Revisited. *Financial Advisor Blog*
<https://leadingadvisor.com/the-four-stages-of-competence-revisited/> accessed 03.06.2022.

- Ritchhart, R. and Perkins, D. N. (2005). Learning to think: The challenges of teaching thinking. In Holyoak, K.J. and Morrison, R.G. (Eds.), *The Cambridge handbook of thinking and reasoning* (775- 802). Cambridge University Press.
- Robinson, C. (2011). Children's rights in student voice projects: where does the power lie? *Education Inquiry*, 2(3), 437-451.
- Robinson, C. (2014). *Children, their voices and their experiences of school: what does the evidence tell us?* Cambridge Primary Review Trust.
- Robinson, C., and Fielding, M. (2007). *Children and their primary schools: Pupils' voices*. Cambridge Primary Review Trust.
- Robinson, C., and Fielding, M. (2010). Children and their primary schools. *The Cambridge primary review research surveys*, 17-48.
- Rodriguez, L.F., and Brown, T.M. (2009). From voice to agency: Guiding principles for participatory action research with youth. *New Directions for Youth Development*, 2009(123), 19-34.
- Rudduck, J. and Fielding, M. (2006). Student voice and the perils of popularity. *Educational Review*, 58(2), 219–31.
- Russell-Mundine, G. (2012). Reflexivity in Indigenous research: Reframing and decolonising research? *Journal of Hospitality and Tourism Management*, 19, e7.
- Saldaña, J. (2013). *The Coding Manual for Qualitative Researchers* (2nd edition). SAGE.
- Sandoval, M. and Messiou, K. (2020). Students as researchers for promoting school improvement and inclusion: a review of studies, *International Journal of Inclusive Education*, 1-16.
- SCAA (1995). *Mandatory Code of Practice for the GCSE*. School Curriculum and Assessment Authority and Curriculum Assessment Authority for Wales. March 1995.
- Schmidt, R. A., and Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological science*, 3(4), 207-218.
- Schnotz, W., and Kürschner, C. (2007). A reconsideration of cognitive load theory. *Educational Psychology Review*, 19, 469–508.
- Scott, D. (1989). HMI reporting of the GCSE. *Journal of Education Policy*, 4(3), 281-287.
- Shaw, S. and Crisp, V. (2011). Tracing the evolution of validity in educational measurement: past issues and contemporary challenges. *Assessment Matters: A Cambridge Assessment publication*, 11, 14-19.
- Shaw, S. and Crisp, V. (2020). An approach to validation: Developing and applying an approach for the validation of general qualifications, *Assessment Matters, A Cambridge Assessment Publication*, Special Issue 3.

- Shelton, S. A., Barnes, M. E., and Flint, M. A. (2019). "You stick up for all kids": (De) Politicizing the enactment of LGBTQ+ teacher ally work. *Teaching and Teacher Education*, 82, 14-23.
- Silverman, D. (1993). *Interpreting Qualitative Data*. London: Sage.
- Sim, J., and Waterfield, J. (2019). Focus group methodology: some ethical challenges, *Quality and Quantity*, 50, 3003-22.
- Smagorinsky, P. (2018). Deconflating the ZPD and instructional scaffolding: Retranslating and reconceiving the zone of proximal development as the zone of next development, *Learning, Culture and Social Interaction*, 16, 70-75.
- Snyder, H. (2019). Literature review as a research methodology: an overview and guidelines. *Journal of Business Research*, 104, 333-339.
- Spalding, V. (2011). 'Is an exam paper greater than the sum of its parts?' *AQA Centre for Education Research and Policy*
https://research.aqa.org.uk/sites/default/files/pdf_upload/CERP-RP-VS-19092009.pdf
accessed 05.07.2020.
- Standish, A., and Perks, D. (2021). The place of public examinations in future school assessment. *Impact*, 12, 19-21.
- Stentiford, L., Koutsouris, G. and Allan, A. (2021). Girls, mental health and academic achievement: a qualitative systematic review, *Educational Review*, DOI: 10.1080/00131911.2021.2007052
- Stern, J. (2018). *What no-one tells you about Bloom's Taxonomy*, [edtosavetheworld blog, 12 June 2018] accessed from <https://edtosavetheworld.com/2018/06/12/what-no-one-tells-you-about-blooms-taxonomy/> .
- Stobart, G. (2008). *Testing times – the uses and abuses of assessment*. Routledge.
- Stobart, G. (2009). Determining validity in national curriculum assessments. *Educational Research*, 51(2), 161-179.
- Stobart, G., and Eggen, T. (2012). High-stakes testing—value, fairness and consequences. *Assessment in Education: Principles, Policy and Practice*, 19(1), 1-6.
- Sun, S., Gao, X., Rahmani, B. D., Bose, P., and Davison, C. (2022). Student voice in assessment and feedback (2011–2022): A systematic review. *Assessment and Evaluation in Higher Education*, 1-16.
- Sweller, J. (2015). In academe, what is learned, and how is it learned? *Current Directions in Psychological Science*, 24(3), 190-194.
- Sweller, J. (2017). Presentation at researchED conference hosted by the Australian College of Educators, Melbourne, 1 August 2017 <https://www.youtube.com/watch?v=gOLPfi9Ls-wandt=1s> .

- Sweller, J., and Cooper, G. A. (1985). The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and instruction*, 2(1), 59-89.
- Sweller, J., and Levine, M. (1982). Effects of goal specificity on means–ends analysis and learning. *Journal of experimental psychology: Learning, memory, and cognition*, 8(5), 463-474.
- Sweller, J., Ayres, P., and Kalyuga, S. (2011). *Cognitive Load Theory. Explorations in the Learning Sciences, Instructional Systems and Performance Technologies*, vol. 1. Springer, New York.
- Sweller, J., van Merriënboer, J. J. G., and Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, 10(3), 251-196.
- Sweller, J., van Merriënboer, J. J., and Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, 31(2), 261-292.
- Tan, H.C.J., Ng, W.L., and Shutler, P.M.E. (2017). Development and field-testing of an instrument for rating cognitive demands of mathematical assessment items. *The Mathematics Educator*, 17(1), 57-78. Retrieved from http://math.nie.edu.sg/ame/matheduc/tme/tmeV17_1/paper3.pdf.
- Taylor, C., and Robinson, C. (2009). Student voice: Theorising power and participation. *Pedagogy, Culture & Society*, 17(2), 161-175.
- Taylor, L. (2005). Washback and impact. *ELT Journal*, 59(2), 154-155.
- Taylor-Egbeyemi, J., Carter, H., and Robin, C. (2023). Thematic analysis of national online narratives on regular asymptomatic testing for Covid-19 in schools in England. *BMC Public Health*, 23(1), 1-13.
- TES (2019). 'Top unis' GCSE demands favour private pupils'. *Times Education Supplement*, 19.08.2019. <https://www.tes.com/magazine/archive/exclusive-top-unis-gcse-demands-favour-private-pupils> accessed 12.06.2022.
- Thomson, P. (2011). Coming to Terms with 'Voice', in: Czerniawski, G. and Kidd, W. (Ed.) *The Student Voice Handbook: Bridging the Academic/Practitioner Divide*. London, Emerald Group Publishing Ltd., 19-30.
- Tikly, L. (2011). Towards a framework for researching the quality of education in low-income countries. *Comparative Education*, 47(1), 1–23.
- Turner, J. (1998). Turns of phrase and routes to learning: The journey metaphor in educational culture. *Intercultural communication studies*, 7, 23-36.
- United Nations (1989). United Nations Charter on the Rights of the Child. <https://www.unicef.org/child-rights-convention/convention-text> accessed 12.06.2022.
- Usher, E. L. and Pajares, F. (2009). Sources of self-efficacy in mathematics: A validation study. *Contemporary Educational Psychology*, 34, 89-101.

- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.
- Waring, M. (2021a). 'Finding your theoretical position.' In Coe, R., Waring, M., Hedges, L.V. and Ashley, L.D. (Eds), *Research methods and methodology in education*, (3rd edition, 15-22). Sage Publications.
- Waring, M. (2021b). 'Grounded Theory.' In Coe, R., Waring, M., Hedges, L.V. and Ashley, L.D. (Eds), *Research methods and methodologies in Education*, (3rd edition, 119-134). Sage Publications.
- Wellcome Trust (2010). '*Marks tell you how you've done...Comments tell you why*' *Attitudes of Children and Parents to Key Stage 2 Science Testing and Assessment*. London: Wellcome Trust.
- Wiliam, D. (1995). It will all end in tiers! *British Journal of Curriculum and Assessment*, 5(3), 21-24.
- Wiliam, D. (1997). Relevance as MacGuffin in mathematics education. In *British Educational Research Association Conference, York, September 1997*.
- Wilkinson, S. (2011). Analysing Focus Group Data. In Silverman, D. (ed.), *Qualitative Research*, 3rd edition. SAGE.
- Williams, J. B. (2006). The place of the closed book, invigilated final examination in a knowledge economy. *Educational Media International*, 43(2), 107-119.
- Willingham, D.T. (2009). *Why don't students like school? A cognitive scientist answers questions about how the mind works and what it means for the classroom*. Jossey-Bass books.
- Winter, G. (2000). A comparative discussion of the notion of 'validity' in qualitative and quantitative research. *Qualitative Report*, 4(3), 1-14.
- Wisby, E. (2011). Student Voice and New Models of Teacher Professionalism, in Czerniawski, G. and Kidd, W. (Ed.) *The Student Voice Handbook: Bridging the Academic/Practitioner Divide*. Emerald Group Publishing Ltd., 31-44.
- Wood, A. (2007). Commentary on Chapter 5 'The Demands of Examination Syllabuses and Question Papers' in Newton, P., Baird, J. -A., Goldstein, H., Patrick, H. and Tymms, P. (Eds). *Techniques for Monitoring the Comparability of Examination Standards*. Qualifications and Curriculum Authority.
- Wood, D., Bruner, J.D. and Ross, G. (1976). 'The role of tutoring in problem solving.' *Journal of Child Psychology and Psychiatry*, 17, 89-100.
- Wood, R. (1991). *Assessment and testing*. Cambridge University Press.
- Wroe, R. (2021). GCSE (9-1) Maths – resources to assist tier entry decision making. Blog for OCR. <https://www.ocr.org.uk/blog/gcse-maths-resources-for-tier-entry-decisions/> accessed 12.06.2022

Zaal, M. and Ayala, J. (2013). 'Why Don't We Learn Like This in School?' One Participatory Action Research Collective's Framework for Developing Policy Thinking, *Journal of Curriculum Theorizing* 29(2), 159-173.

This page has been left blank intentionally

Appendices

Appendix A: Pilot Study – Information Sheet, Consent Form and Questionnaire

Note: the pilot study sample was divided into two groups. These were initially called Group 1 and Group 2. In the thesis, the names of these groups were changed, to Group P and Group Q. Group P made their predictions of 'difficulty' first, then attempted the questions. Group Q answered the questions first, and then reported their estimates of question 'difficulty.' Group P's questionnaire (Group 1) is given here.

Difficulty in GCSE Maths questions: Group 1

You are invited to take part in a study that I am conducting as part of my PhD research at Durham University. I want to find out more about students' views of what makes GCSE maths questions more or less difficult. Your opinions are important to me.

My name is Mr Fowler. I am the researcher, and I am also the Principal at Lord Lawson. I am carrying out this research to help improve understanding of teaching and learning. My email address is a.t.fowler@durham.ac.uk and you are welcome to ask me any questions about this research.

This study has received approval from Durham University School of Education. My supervisor is Dr Dimitra Kokotsaki and her email address is Dimitra.kokotsaki@durham.ac.uk (<mailto:Dimitra.kokotsaki@durham.ac.uk>).

Please read these statements and show that you understand them and agree with them.
Then please answer the questions that follow.
This questionnaire will take around 20-30 minutes.

* Required

1. Purpose of the study: This study aims to find out from students how you understand the different ways of making GCSE maths questions easier or more difficult. Understanding this will help you and your teachers prepare better for exams. You have been invited to take part because you are studying GCSE maths. Your views are important.

I have read this statement and I understand the purpose of the research in which I am taking part. I have been given the opportunity to ask questions via email about the study. *

Yes

No

2. Your participation is voluntary: you do not have to take part. If you do agree to take part, you can withdraw at any time; you do not need to give a reason. If you have any questions, you can email me or my supervisor on the addresses above. If you agree to take part, you will complete two online questionnaires and answer some GCSE maths questions.

You will be asked about how confident you feel about maths, and you will be asked to estimate how difficult you think the questions are. These questions are like ones you usually study at school. There are no risks to you in taking part.

In a later part of the study, I may ask you if you would like to take part in a short discussion about your answers. This is also voluntary: you do not have to take part.

I understand that my participation is voluntary and that I am free to withdraw at any time without giving a reason. *

Yes

No

3. Your answers to the questions will be kept confidential. No personal data from this study will be shared or published at any time. If I use any part of your answers in my study, and if my study is published, your answers will be anonymous, and no-one will be able to identify you from the study.

I will give some feedback to you and your teachers, around one month after the completion of the questionnaires. I will write up this study as a PhD thesis, and this will be kept by Durham University in print and online. Other researchers may have access to my study, for research purposes only.

Once the study is completed, your anonymous answers will be destroyed.

I understand how the data will be stored and what will happen to the data at the end of the project. *

Yes

No

4. I understand that my words may be quoted in the write-up of this study, and in other research publications, but that I will not be identified. *

Yes

No

5. I agree to take part in the project. *

Yes

No

6. Your name [first name and surname] *

7. Your school: *

8. The gender you identify with: *

Female

Male

Non-binary/Prefer not to say

9. I am interested in learning about what you think makes maths questions easier or more difficult.

All these questions have been taken from mathematics GCSE past papers. Read each question, then estimate how easy or difficult you think it is, and give some reasons.

You do not need to work out the answer to the question, but thinking through how to answer it may help you decide how difficult it is.

Later in the questionnaire, you will be asked to answer the question.

Question A:

Reuben hires a car.

It costs £150, plus 85p for each mile he travels.

When Reuben hires the car, its mileage is 27,612 miles.

When Reuben returns the car, its mileage is 28,361 miles.

How much did Reuben pay to hire the car? (4 marks) *

- Very Easy
- Easy
- Neither easy nor difficult
- Difficult
- Very difficult

10. Explain why you think Question A is easy or difficult. *

11. Question B:

Finn has two bags of counters.

He takes a counter at random from each bag.

The probability that he takes a red counter from the first bag is 0.3

The probability that he takes a red counter from the second bag is 0.4

What is the probability that he takes at least one red counter? (4 marks) *

- Very easy
- Easy
- Neither easy nor difficult
- Difficult
- Very difficult

12. Explain why you think Question B is easy or difficult *

13. Question C:

60% of the people in a town are males.
20% of the males are left-handed.
21.6% of all the people are left-handed.

Work out the percentage of the people who are not male who are left-handed. (5 marks) *

- Very easy
- Easy
- Neither easy nor difficult
- Difficult
- Very difficult

14. Explain why you think Question C is easy or difficult *

15. Question D:

The value of a car, £V, is given by

$$V = 16,500 \times 0.82^n$$

where n is the number of years after it is bought from new.

a) Write down the value of the car when new (1 mark)

b) Write down the annual percentage decrease in the value of the car (1 mark)

c) Show that the value of the car after 4 years is less than half its value when new (2 marks) *

- Very easy
- Easy
- Neither easy nor difficult
- Difficult
- Very difficult

16. Explain why you think Question D is easy or difficult *

17. Question E:

Solve

a) $4x = 56$ (1 mark)

b) $8x - 6 = 46$ (2 marks)

Solve by factorising

c) $x^2 + 11x + 30 = 0$ (3 marks) *

Very easy

Easy

Neither easy nor difficult

Difficult

Very difficult

18. Explain why you think Question E is easy or difficult *

19. Question F:

A circle has radius 6cm

Calculate its circumference

Give your answer in centimetres, correct to 1 decimal place (3 marks) *

- Very easy
- Easy
- Neither easy nor difficult
- Difficult
- Very difficult

20. Explain why you think Question F is easy or difficult *

21. In this second part of the questionnaire, you are asked to work out the answers to the questions you have already seen.
You may find that it helps you to have a spare piece of paper on which you can do your workings.

Question A:

Reuben hires a car.

It costs £150, plus 85p for each mile he travels.

When Reuben hires the car, its mileage is 27,612 miles.

When Reuben returns the car, its mileage is 28,361 miles.

How much did Reuben pay to hire the car? (4 marks) *

22. Question B:

Finn has two bags of counters.

He takes a counter at random from each bag.

The probability that he takes a red counter from the first bag is 0.3

The probability that he takes a red counter from the second bag is 0.4

What is the probability that he takes at least one red counter? (4 marks) *

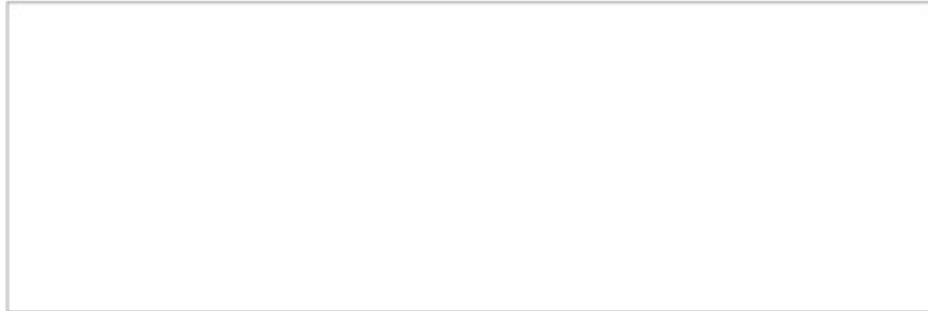
23. Question C:

60% of the people in a town are males.

20% of the males are left-handed.

21.6% of all the people are left-handed.

Work out the percentage of the people who are not male who are left-handed. (5 marks) *



24. Question D:

The value of a car, £V, is given by

$$V = 16,500 \times 0.82^n$$


where n is the number of years after it is bought from new.

a) Write down the value of the car when new (1 mark)

b) Write down the annual percentage decrease in the value of the car (1 mark)

c) Show that the value of the car after 4 years is less than half its value when new (2 marks)

Remember to give answers to all three parts of the question - a) b) and c) *



25. Question E:

Solve

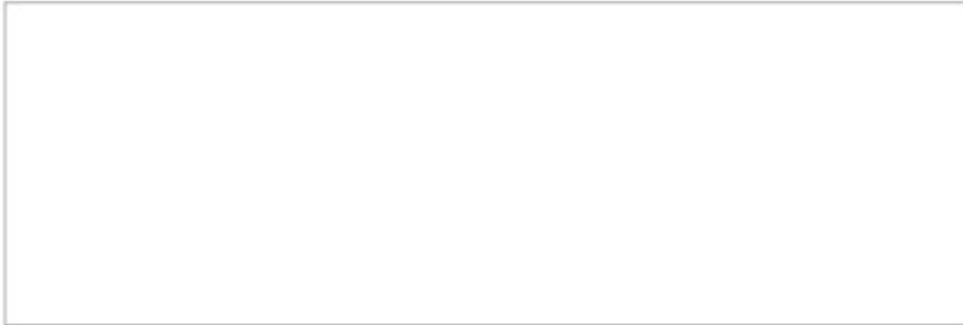
a) $4x = 56$ (1 mark)

b) $8x - 6 = 46$ (2 marks)

Solve by factorising

c) $x^2 + 11x + 30 = 0$ (3 marks)

Remember to give answers to all three parts of the question - a) b) and c) *

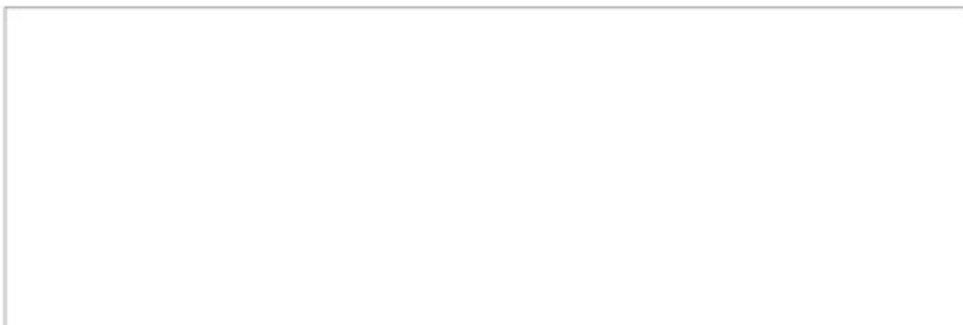


26. Question F:

A circle has radius 6cm

Calculate its circumference

Give your answer in centimetres, correct to 1 decimal place (3 marks) *



Difficulty in GCSE Maths questions

September 2022

You are invited to take part in a study that I am conducting as part of my PhD research at Durham University. I want to find out more about students' views of what makes GCSE maths questions more or less difficult. Your opinions are important to me.

My name is Mr Fowler. I am the researcher, and I am also the Principal at Lord Lawson. I am carrying out this research to help improve understanding of teaching and learning. My email address is a.t.fowler@durham.ac.uk and you are welcome to ask me any question about this research.

This study has received approval from Durham University School of Education. My supervisor is Dr Dimitra Kokotsaki and her email address is Dimitra.kokotsaki@durham.ac.uk

Please read these statements and show that you understand them and agree with them. Then please answer the questions that follow.

The questions and questionnaire will take around 40-50 minutes.

1.	<p>Purpose of the study. This study aims to find out from students how you understand the different ways of making GCSE maths questions easier or more difficult. Understanding this will help you and your teachers prepare better for exams. You have been invited to take part because you are studying GCSE maths. Your views are important.</p> <p>I have read this statement and I understand the purpose of the research in which I am taking part. I have been given the opportunity to ask questions via email and in person about the study.</p> <p>Yes <input type="checkbox"/></p> <p>No <input type="checkbox"/></p>
----	--

2.	<p>As part of your usual maths lessons, you will answer some past paper maths questions and you will complete a questionnaire.</p> <p>You will be asked to estimate how difficult you think the questions are, and you will be asked about factors that, in your view, make questions more or less difficult. These questions are like ones you usually study at school. There are no risks to you in taking part in this study.</p> <p>Your participation in this research is voluntary. You do not have to take part. If you do agree to take part, you can withdraw at any time, and you do not need to give a reason. In a later part of the study, I may ask if you would like to take part in a short discussion about your answers. This is also voluntary: you do not have to take part.</p> <p>I understand that my participation is voluntary and that I am free to withdraw at any time without giving a reason.</p> <p>Yes <input type="checkbox"/></p> <p>No <input type="checkbox"/></p>
3.	<p>Your answers to the questions will be kept confidential. No personal data from this study will be shared or published at any time. If I use any part of your answers in my study, and if my study is published, your answers will be anonymous, and no-one will be able to identify you from the study.</p> <p>I will give some feedback to you and your teachers, around one month after the completion of the questionnaires.</p> <p>I will write up this study as a PhD thesis, and this will be kept at Durham University in print and online. Other researchers may have access to my study, for research purposes only.</p> <p>Once the study is completed, your anonymous answers will be destroyed.</p> <p>I understand how the data will be stored and what will happen to the data at the end of the project.</p> <p>Yes <input type="checkbox"/></p> <p>No <input type="checkbox"/></p>

4.	I understand that my words may be quoted in the write-up of this study, and in other research publications, but that I will not be identified. Yes <input type="checkbox"/> No <input type="checkbox"/>
5.	I agree to take part in the project Yes <input type="checkbox"/> No <input type="checkbox"/>
6.	Your name [first name and surname] Your signature:
7.	The gender you identify with: Female <input type="checkbox"/> Male <input type="checkbox"/> Non-binary/prefer not to say <input type="checkbox"/>

This page has been left intentionally blank

Difficulty in GCSE Maths questions

You may use a calculator for these questions. Please show your workings.

Assume the value of π to be 3.14

1. Reuben hires a car.
It costs £150, plus 85p for each mile he travels

When Reuben hire the car, its mileage is 27,612 miles

When Reuben returns the car, its mileage is 28,361 miles

How much did Reuben pay to hire the car?

(4 marks)

2. Finn has two bags of counters.
He takes a counter at random from each bag.

The probability that he takes a red counter from the first bag is 0.3

The probability that he takes a red counter from the second bag is 0.4

What is the probability that he takes at least one red counter?

(4 marks)

3. 60% of the people in a town are males.
20% of the males are left-handed.
21.6% of all the people are left-handed.

Work out the percentage of the people who are not male who are left-handed.

(5 marks)

4. The value of a car, £V, is given by

$$V = 16,500 \times 0.82^n$$

where n is the number of years after it is bought from new.

- a) Write down the value of the car when new

(1 mark)

- b) Write down the annual percentage decrease in the value of the car

(1 mark)

- c) Show that the value of the car after 4 years is less than half of its value when new.

(2 marks)

5. Solve

a) $4x = 56$

(1 mark)

b) $8x - 6 = 46$

(2 marks)

c) Solve by factorising

$$x^2 + 11x + 30 = 0$$

(3 marks)

6. A circle has radius 6cm

Calculate its circumference

Give your answer in centimetres, correct to 1 decimal place

(3 marks)

7. A bag contains 12 counters.
6 are red, 4 are blue and 2 are yellow.
A counter is taken from the bag at random.

What is the probability that the counter is

a) Red

(1 mark)

b) Yellow

(1 mark)

c) Green

(1 mark)

Appendix D: Main Study – Questionnaire

Difficulty in GCSE Maths questions

Please read the whole paper carefully before answering these questions

a) Considering all the parts of each question together, how difficult do you think these questions are? Please circle your answer

Q1 “Reuben hires a car...”

Very easy	Easy	Neither easy nor difficult	Difficult	Very difficult
-----------	------	-------------------------------	-----------	----------------

Q2 “Finn has two bags of counters...”

Very easy	Easy	Neither easy nor difficult	Difficult	Very difficult
-----------	------	-------------------------------	-----------	----------------

Q3 “60% of the people in a town are males. 20% of the males are left-handed...”

Very easy	Easy	Neither easy nor difficult	Difficult	Very difficult
-----------	------	-------------------------------	-----------	----------------

Q4 “The value of a car, £V, is given by $V = 16,500 \times 0.82^n$...”

Very easy	Easy	Neither easy nor difficult	Difficult	Very difficult
-----------	------	-------------------------------	-----------	----------------

Q5 “Solve $4x = 56$; $8x - 6 = 46$; solve by factorising $x^2 + 11x + 30 = 0$ ”

Very easy	Easy	Neither easy nor difficult	Difficult	Very difficult
-----------	------	-------------------------------	-----------	----------------

Q6 “A circle has radius 6cm. Calculate its circumference”

Very easy	Easy	Neither easy nor difficult	Difficult	Very difficult
-----------	------	-------------------------------	-----------	----------------

Q7 “A bag contains 12 counters. 6 are red, 4 are blue, 2 are yellow”

Very easy	Easy	Neither easy nor difficult	Difficult	Very difficult
-----------	------	-------------------------------	-----------	----------------

b) Which question did you rate as the **least** difficult?
Please give your reasons

c) Which question did you rate as the **most** difficult?
Please give your reasons

d) Questions 2 and 7 are both about probability.
Which one do you think is more difficult?

Please explain why you think this is

e) What factors do you think make examination questions in Mathematics more difficult?

f) Do you think that GCSE examination questions in Mathematics correctly distinguish between more able and less able students? Please circle your answer

Yes, completely	Yes, to some extent	Unsure	No, not completely	No, not at all
--------------------	------------------------	--------	-----------------------	----------------

Please explain your view

Your name:

Appendix E: Focus Group 1 Transcript

Monday 19 December, 2022, 11:30am

Year 11 Top Set Maths Group

Researcher [AF]; Students R31M, R34M, R36M, R49M and R51M.

[Introduction and explanation given, including purpose of the focus group and privacy notice.]

AF: I've got a few questions for you. First a sort of general one. When you look at an examination question in maths, I wonder, what do you look at first? So, say you've got a question, like a past paper question, something you might be doing in lessons, what do you look at first? What's your focus? What do you go for? Maybe, maybe you all do the same thing, maybe you all do different things. Erm, actually I forgot to ask for names, so we've got [student R49M], on my left, and then [student R34M], and then [student R36M], and then [student R31M] and [student R51M]? Did I get that right? Thank goodness for that! Yes, OK, would somebody like to start off? Er, when you look at an examination question, what do you look at first? [Student R34M], can I ask you that one?

R34M: Erm, usually the marks, like how much it's worth, and you see how long you can spend on it.

AF: OK. Does anybody else look for that, or do you look for different things?

R31M: Er, I tend to look for just the very end of the question, just exactly what it's asking for.

AF: That's [student R31M] answering...

R31M: ...yes, and then I can just find the important bits of information in the question, like the numbers and that, instead of reading the whole question.

AF: OK. Talk me more through that approach.

R31M: So it's like, "How much did Reuben pay to hire the car?" So I'll look at how many miles he's done and I'll look at how much it costs, and so starting from the start of the question and, erm, reading it there.

AF: OK. Thank you. What about you, [student R51M]?

R51M: Erm, I usually, like, look through all the numbers first, and I'll see what the question's actually asking, and then I'll... go back to the numbers and then work out important, like, differences between them and everything like that.

AF: OK. Thank you. What about you, [student R49M]?

R49M: I usually just look at the numbers first, to be honest, and sort of figure out what I need to do from there. Cos... I just quite like numbers, to be fair.

AF: Huhu. So you go for the numbers...

R49M: ...yeah...

AF: ...first...

R49M: ...yeah.

AF: Yeah. [Student R36M]?

R36M: I, I'm like [student R34M], I'll look at the marks as well, cos it, like, tells us how much work I need to do, cos there's a difference, obviously, between like a 2-mark question and a 5-mark question. So, straightaway I can see if I need to work out more to do with the numbers, or if I've missed out a step and it's more marks than, like, a 2-marker.

AF: OK. Erm. Right. [Student R34M], you've got more to add there? No? You look like... No. And so... so, once you'd sort of like basically decoded the question there, what do you do next? What would you do next, [student R49M]?

R49M: I think after I've checked out all the numbers and figured out what I need to do, erm... I'd sort of... I'd do my first step, which is sort of like... Like for the example of the first one, I think I'd figure out the difference between the two mile numbers. The first, obviously, because you need that to be able to, like, figure out the answer to the question, and it's like a very key part of the question, so I'd make sure I'd answer the key parts, like, very obviously, so it's easier to get marks.

AF: Uh-huh. So you'd make sure you'd done that key part of the question...

R49M: ...Yeah...

AF: ...First, uhuh... And would you – I'm interested in anybody's response here, would you always use the same approach – looking at the length of it, or looking at the numbers first – would you always use the same approach, or does that vary from, from question to question?

R36M: Erm, this is [student R36M], erm... I think, if it's, if it's like a longer question and multiple lines and, like, names and things like that, I would look at the marks first, cos that would tell us; but if it's a short, like, "work out", like two fractions for example, then I could just see what I need to do, I could just use the numbers straight away, without looking at the marks, cos I would just, I would know what to do straight away.

AF: OK, so you would just go straight, straight into the question at that point?

R36M: Yeah.

AF: OK. Erm, and I'm interested, also, if you're approaching a question in an exam, how is that different, if at all, from how you'd approach it in class? What do you think, [student R31M]?

R31M: I would say... I answer them quite similarly. What I always do is, I try and make an equation and go from there. So I'd see what the question's asking and make an equation based on that, add any unknowns that are needed. And I'd start ignoring it

being like a question as much and I would just do the maths for it. Erm, in class I would do something quite similar, but I might like draw in a graph as well instead.

AF: OK. What, what about you [student R51M]? Would you approach differ if it was in an exam or in class?

R51M: If it was in class I'd probably just try and do it mentally, but whereas in a test I have to... [unclear]

AF: Can you say that really clearly, for the recording?

R51M: If I'm, like, in a test, I'd show my working out, whereas in class I'd usually do it mentally in my head.

AF: All right, so there might be some bits that you didn't write down?

R51M: Yes.

AF: OK. Good. Thank you. Right, now, looking at Question 1, erm, [student R49M], you've already given us a start, really, about how you'd solve question 1. Anybody else like to tell us about what you'd do, how you'd approach or solve that question number 1?

R31M: Yeah, sure. It's [student R31M]. Erm, what I would do is, I would see how many miles he's travelled, cos you've got to work out how many miles for each 85p additionally, so you'd do 28,361 take away 27,612, erm, and you'd just times that by 85 pence, and then add that to the 150.

AF: OK... Well, this is a question that most people answered correctly. Does anybody disagree with [student R31M]'s, or have a different method? No? There's lots of people shaking their heads there. OK, so... I think we're all fairly clear on that one then. Let's turn now to question 3.

R31M: I didn't like this question!

AF: [laughs] Now, [student R31M], you've just told me you didn't like this question. Erm, can you tell me a bit more about what you didn't like about it?

R31M: The main reason I didn't like this question – we've just had a test recently and I didn't like it for the same reason – is the wording of it. If there's a question where it's mainly numbers, it's quite easy to get it correct, but if it's, like, "work out the percentage of people who are not male who are left-handed" if you try and do it over the whole population per se you just miss out the left-handed bit at the end, just the wording of it can throw off all your maths. Erm, in one of the tests we had, it was, like, "what is an assumption that could be made about it?" and it's like, you could make many different assumptions and it's like, which one could be right, or could be wrong. Erm, so I think when it comes down to not being maths any more, that's when it gets tricky.

AF: Now, [student R31M] said something interesting there, about "not being maths any more." Anybody want to pick up on, on that? [Student R34M]?

R34M: It's like, it gets, cos like when the question's worded badly, it gets to the point where it's more about you being able to read instead of being able to do the maths; you have to be able to interpret what it's saying, before actually being able to do what it's asking.

AF: Tell me more about that – it's really interesting.

R34M: Like, for like the first one, it's really clear: it says, "How much did Reuben pay to hire the car?" You know what you need to do – you just need to do the maths. But if the words are a bit 'off', you have to interpret what it's asking before you can even start with the maths. If that's not your strong point, you've kind of failed already.

R31M: Could I say something as well?

AF: Yup...

R31M: ...it's [student R31M] again. There was a Further Maths mock we had, towards the end of Year 10, and the final question of it, erm, wasn't written out quite well: I wasn't able to tell whether you were times-ing it by a fraction, whether it was to the power of a fraction or if it was a mixed number. And, because of that, I wasn't able to tell what it was asking to do, so I got the question wrong.

AF: Right, so you got the question wrong because you weren't able to tell...

R31M: ...because the question was not formatted well and I couldn't ask, I couldn't tell what it was asking me to do.

AF: Uh-huh. [Student R51M], you're just, sort of nodding along with that. D'you want to add to anything there?

R51M: I feel, like, lot of the time maths questions can be like, worded quite, erm, particularly like they want you to do a specific thing, like it's just over-complicating the question. And it just makes it, like, more longwinded to do.

R36M: Yeah. This is [student R36M]. Back to [student R34M]'s point, I mean, I think, like, I think a lot of people could do this question 3 if it was put out easily, cos the maths of it isn't very hard. But the like, the last line "who are not male who are left-handed" it can just really throw you off and, like, what are you trying to work out from the question. And I think that's why more people would get it wrong, and it's not just the maths, which is not probably what the exam board is probably looking for. It's like... the question... if they're wanting to see how good you are at maths cos it's a maths paper, if you read, if you cannot read it properly or if you misinterpret the question, then I don't think that's what it should be assessed, rather than it just being the pure maths skills.

AF: OK. What about if, if actually they're trying to assess something as well as the maths skills? If they're trying to assess, er, your problem-solving abilities? Anybody have a... a comment on that?

R31M: Erm, this is [student R31M]. I would say – yes, they probably are looking for problem solving and that – but when they test the people and they all get the same wrong

answer because they've not got the wording of the question correct, I feel like that would be because the question's worded badly. If everyone got different answers, it could be like they've all solved it slightly differently – erm, but if they get the same wrong answer – if everyone puts the percentage of the whole population instead of one thing... erm, dunno.

AF: OK. Yeah, no, it's interesting, an interesting answer. Yeah. [Student R49M], anything to add on that?

R49M: I think personally with me it's like, erm... words just make it a little, a lot harder and problem solving sort of borders maths and completely different like skills; whereas I think realistically you shouldn't really need to like, have specific reading skills, where you can understand what they're saying, to be able to do the maths question – you should just be able to just answer it without having to sort of like read lower into it – cos you don't, you shouldn't have to interpret anything in maths: because it's just sort of like "this is this", it shouldn't be like "this could be this or this", you know, I mean...

AF: So you think it should be – I don't want to put words in your mouth – can you tell me a bit more about what you mean there?

R49M: I think I just mean it should be more like, sort of specific towards what you need to answer it with, instead of sort of, meaning you need to, like, sort of, think of what it has to be...

AF: ...Hmm...

R49M: It should, like, tell you what it needs.

AF: So looking at that question, can anybody talk me through how you developed your thinking for answering this question? What sort of strategies did you use? Cos there are lots of different ways you could do it. Anybody talk me through how you might approach this? [pause] [Student R34M], what would you do to solve this question?

R34M: I think – I don't know if I got it right, to be honest – what I think I'd try to do was work out all the people in the town who are left-handed, and then, like, obviously work out... it gives you how many males there is, and then I try to do some weird thing – like, I don't know if I did a ratio or something, where I did, like, the number of men, the number of left-handed and tried to work it out from there.

AF: OK. So [student R34M] used a ratio method. Did anybody else, erm, would anybody else approach this through a ratio? So, there's a few heads being shaken... So, [student R51M], you're shaking your head – how would you approach this? What, what method would you use?

R51M: I think I can remember – to begin with I'd times the percentages by a hundred and then put them into a tree diagram, so I could, like, see what was being said. So, for example, I'd put 600, like, on top and then a line being drawn down to females and then on the line to males I'd put like 400 right-handed males and 200 left-handed males – and then continue it like that.

AF: OK, so you used a tree diagram to, to help with that. [Student R31M], you're nodding along with that.

R31M: I didn't do it as a tree diagram, but I feel like that would have made a lot more sense than what I did. I think I tried to do, for instance, 60% of the people of the town are males, 20% are left-handed, which are males, so I times'd them together to get, like, 12%, and then I would take that off of the 21.6% to get whatever was left, which had to be females who are left-handed... erm. But that gave me the wrong answer to the question, because it's asking for "who are left-handed" but of that "who are not male". So, instead of females who are left-handed of the whole population, it wanted people who are left-handed out of the females. So I got the wrong answer to that.

AF: Uh-huh. OK. [Student R36M], what was your, how would you approach this?

R36M: I think I put it into three separate tables of males, females and the whole population. And then have a right-handed and a left-handed, and then try to work out a value for each place in the table. And then, for the question "who are not male who are left-handed" I think I looked at how many overall were left-handed and then saw, and then just took my value from the table which is easy enough to find, to work out if you've got the three values that you're given, and then you could just, like, insert them into a fraction and then turn the fraction into a percentage to get your answer.

AF: OK. So you used a table?

R36M: Yes.

AF: What about you, [student R49M]? How would you approach this?

R49M: I did it exactly the same as [student R31M] – I got it completely wrong, cos I did, erm... I got the 12% from multiplying the males and the left-handed males together and just took that away from the 21.6 and left that as my answer, I'm pretty sure.

AF: OK. Right. Thank you, thank you very much. Looking at questions 4 and 5 now, [coughs] excuse me. So, question 4 is one these kind of 'wordy' questions and you know you've already talked a bit about words questions versus, um, numbers questions. And question 5 is much shorter, just numbers, isn't it? So, I wonder how your approach would differ from question 4, the 'wordy' one, to question 5, the one that's got just straight numbers in it? And which, which questions would you find easier to understand, and which ones would you find easier to solve, and whether you could explain that? I'm also interested in why some other people might disagree. So, I'm asking you some quite complicated things here: how would you approach one set of questions versus another set, erm, which type d'you find easier to understand, and could you think that other people might disagree with that? OK, who wants, who wants to start with that?

R31M: I can.

AF: OK, [student R31M]?

R31M: Yeh. I think question 5 I would say would be more understandable to me, because I don't need to worry about reading any words – so maybe not more understandable but I get them quicker cos I would just immediately see, well for instance, I would need to divide this by 4, I need to add 6 and divide it by 8. But for the ones on question 4, each question just has a single line, which is quite to the point. Erm, I don't know if it's also because we've practised these types of questions lots in lessons – um, but we've done them, like, to death, so we know how to do these ones.

AF: These ones?

R31M: On question 4, sorry, with the, um, the words, "Write down the value of the car when it's new" erm, because, we've gone over the fact that in the equation V equals whatever times whatever to the power of n , the first one is what it would be as new. Erm, I think part (c) of question 4, some people could struggle with, um, the showing that after 4 years, because you've got to do like two parts of it – you've got to work out what half of the value would be and what the other one would be. Um, but I think, overall, everyone would find question 5 easier.

AF: OK. Thank you. [Student R36M]?

R36M: This is [student R36M]. I think, me myself, I would agree with [student R31M], seeing the 5 as easier cos everything's straight to the point; however, I think lots of people might struggle with 5 more because there's less to go off, and there's less to, like – if you don't know what's going on you won't have a clue because all it says is, "Solve" for number 5(a), so if you don't know really what's going on with the question you've got no clue and, like, having a go at it or anything, but on 4, you would, er, you've got lots of words to help understand, you've got context given at the start of 4 before even any questions start, so it might be easier to get an answer if you're not quite sure about, if you're like not quite sure of what to do.

AF: Thank you. [Student R34M]?

R34M: Er, I feel like, if you just do like what we talked about at the start and look at the marks, then you can kind of group them in the same, cos in 4 you can look at it and go, like, "ah, there's loads of words," but if you look and say, "ah, there's just 1 mark," it just takes like all the intimidation out of the question. And it'll, you can look at it basically in the same way as you look at question 5 as they're both just simple 1-mark questions. And it'll, it's not going to be difficult to do just one step, and you can just work from there.

AF: OK. That's a, that's a really interesting insight there – can you, can you tell me a bit more about that?

R34M: Well...

AF: ...I like the way you said it "takes the intimidation out of the question"...

R34M: ...Yeah. Cos, a lot of people, like, they might look at question 4 and they'll see, oh, it's a maths question but they'll see like 6 lines of words, but seeing it's one mark it'll make people think it cannot be hard, it's not going to be difficult, so it'll calm them down,

maybe make them just read through. And then the fact that it's split into 3 parts, it'll also, like – it breaks it down. So you can look at the two questions the same way, cos they're worth the same amount, they're just worded slightly differently.

AF: OK. Thank you. Anyone else got anything to add to that?

R31M: Erm, yeah, a bit. It's [student R31M]. I think what [student R34M]'s saying about the intimidation is definitely true, cos if you get, like at the very end of a paper, it might be a question that has just like a line of words or 2 lines of words and it might be, like, 6 marks, I feel like those questions could be really intimidating. I feel like, in questions where there are lots of marks, the more information given, word-wise, even if it makes it more open to interpretation, does make it more accessible to lots of people, cos they'll be like, right, well, towards the start of it's going to be the first couple of marks, towards the end's going to be the last couple of marks and they do manage to, like, work through it.

AF: OK, so they develop a strategy to work through that. OK. [clears throat] Thank you very much. Now, I'm interested in how students, or how you think that students like you improve your skill in mathematics – you know, you're all in the top set, you're pretty good at maths. So, erm, how did you, how did you improve your skills, how do you still improve your skills, can you talk to me about this, and how did you develop those skills? [Student R51M], can I start with you?

R51M: Um, well, I see. When I started doing like maths in primary school I always liked, we used to love playing this game called Kings and Queens, where we would like practise our times tables and ever since then I've always like loved doing maths, like, I just love anything that involves like working out numbers and anything like that, because it has a definite answer, whereas in things like science or like English it doesn't always have a definite answer within it, so I like it for that reason.

AF: OK, so that explains why you, why you like maths. Yeah. But how did you develop your skills? How did you become better at it, d'you think?

R51M: I feel like, once I got confident with like the key basic parts, like, times tables and everything like that, it really helps your whole maths, like come together, in terms of, like, I can look at question 5 now and just divide it by 4, which is 14, then just, because I know that that's the times tables for it. It's just basic – once you have the basics it's not as complicated as what it seems.

AF: Thank you. [Student R49M], can you talk me through how you developed your skills?

R49M: I think just sort of like learning new things was the best for me. Like, especially with like maths, just sort of doing things you've never done before, learning completely different things from what you were used to sort of helped me, because I was always, like, I didn't really, know what I mean, I didn't sort of... erm, thinking of the word... like, if I don't know something and then I know it I'm obviously getting better at something. So if I'm learning new things then I'm constantly getting better at it, which means if I just keep looking at things I don't know then eventually I'll learn them and then I can get better at maths.

- AF: OK. [pause] Anyone else want to talk to me about how you develop your skills in maths?
- R34M: It's [student R34M]. I feel that either just, like, learning or teach yourself like a couple of different methods to do the same question is probably the best way, because then you can choose one that works best for you. So there might be like question 4 – I could do something completely different to [student R36M] and get the same answer and it's just about working and finding a way that works for you to answer the question. Cos [student R31M] does crazy little equation things for stuff and it just doesn't make any sense at all! [laughs]
- AF: Doesn't make any sense to you?
- R34M: Yes! But it obviously like makes sense to him, cos he's like a genius child. [laughter] So it's just about like working out which method works best for you as yourself.
- AF: Right? [Student R31M], do you want to come in there? D'you think you do learn maths in a different way from other people?
- R31M: I think to a degree, yes. Erm, obviously, like [student R51M] said with the basics, you need to know the basics if you're ever going to get anywhere. Erm, but if you've got like a couple of tricks up your sleeve, like, I like differentiation, which is a fancy way of saying that I'm able to find the turning point of a graph without doing, er, completing the square, and that helps me because I actually don't understand it at all, like lots of the people in our class use it, but I don't understand it. And if I use like a higher method like differentiation, I'm able to get the question. Erm, but I think the main way that you can learn and get better is to do questions and see where you go wrong. Like I used to get negatives wrong all the time, so by seeing that I'm able to check, have I got everything correct when it comes to that. I think mainly just practice and doing questions that you wouldn't usually be comfortable with... erm, questions that will be on a test that you usually wouldn't do, because when you see them you'll actually be able to do them.
- AF: Uh-hmm. OK. What about you, [student R36M]
- R36M: Um, you've got... When an exam board makes a maths paper they're always going to have, like, a set criteria that they've got to meet for it, and they're going to have, like, a word question, they're going to have like easier questions like number 5 where it's just they're stated what you need to do, so I think you need to work on the skills of the question – not particularly on a certain topic, like, anything, but working out how, like, if you get a word question, how you would tackle that straight away, and like, how would you go about, like, that question compared to that question because of the words and not in particular like the maths skills.
- R31M: Yeah, I agree with you there.
- AF: So it's about, sort of, maths techniques, is it? On cracking different questions?
- R31M: Learning how to answer questions, as opposed to actually what the questions might be. Erm, so yeah. Learning how to answer a question, such as question 4, which we've

gone over in class lots, so we could apply that to any variation of that question because we've learned the skill for it.

AF: Tell me more about that.

R31M: Erm, so for instance, this question is about the value of the car when it's new, but if another question came along that was asking what's the value after 3 years, so because we've done the ins and outs of this question, we've learned how to answer it, we'd still be able to do it, even though it's like a completely different question. Erm... if it's asking you to "solve" something, we know we need to find the value for it; if it's asking us to "prove" something, we know we'll have to use algebra to show that without a shadow of a doubt something is something. So instead of "showing" it, like, using an example of, if you say that an even number plus an odd number will always be odd, it's different showing it, like $3 + 2$, as it would be to say, like $n + (n+1)$. It's just different ways of answering them: once you learn those, no matter what the question is, you'll have a good shot at it.

AF: That's really interesting; thank you. Finally, how do you think teachers can make maths more interesting and understandable for students? [pause] Maybe they can't...

R31M: Erm, this is [student R31M]. I would say that in our school we've got a quite good maths department, a really good one actually, and they're able to engage everyone quite well. Erm, I think the ways that they do that is by going through with the class together and then helping students individually that are struggling. But also to give those that are further on challenges, so that they can keep interested, instead of doing the same work. Erm, I think the main way that they could keep students interested is by having that level of individuality, cos the way that I do something would be different to the way that [student R34M] does something, so if the teachers are able to give us each work that works for that, it would be quite good.

AF: OK. [Student R34M], you got a name check there – what do you think?

R34M: I kind of agree with [student R31M]. It's like about variety in what you're doing, so that, say there's a kid who doesn't understand maths and doesn't enjoy it, and they just do the same lesson, basically, for two years straight, they're just... they're going to lose interest, then once you've lost interest, you're not going to want to really get back into it, and you're never going to enjoy it again. So, if they keep changing the lessons, they might find something that engages, like, a struggling student, and that might help them later on, like, cos they want to learn more... (I'm not really sure, to be honest.)

AF: What d'you think, [student R49M]?

R49M: Personally, I think you do have to sort of push people to their limits when it comes to maths. Cos if you give someone something that's not going to challenge them, they go, oh that's quite boring and they're not going to enjoy it, but if you give someone something that's at the top of their ability, that they can, like, they can learn how to do but they don't already know how to do, that'll challenge them to do something, like, harder than they're actually doing. So if you actually, like, sort of, push people... out of their comfort zone when it comes to answering questions, you'll be, you'll be more

likely to learn if you're answering harder questions, whereas if you're answering questions that you can answer with ease then there's no point in answering them because if you're just given that question over and over again you're always going to get the same answer, so if you get, different, like, harder questions, you might spend more time on them and it'll help you learn as a whole, because if you just get a question that takes you, like, two second to do you're not really gonna learn much.

AF: OK. What about a... Can a question be too hard?

R49M: I feel like... a question can be too hard, but I feel like, sort of, when you have sort of limits, you're able to, like, limit yourself, as well as, sort of, that answer... I've lost myself.

AF: Can anybody help him out? I'm really interested in what you're saying there, [student R49M].

R49M: I'll figure it out.

AF: So, yes, this idea about... So [student R49M]'s told us about you get better by doing things which are out of your comfort zone, a bit too hard, but my question is, can you, can you have a question that's too hard, and if you do have a question that's too hard, what effect does that have?

R36M: Yeah... This is [student R36M]. I think if you've learned... if you've got so, if you've done a topic and you've done the content for that topic, if you go to the higher end of the spectrum for that, like the difficulty of questioning for that topic, I don't think it could be too hard, because you're, you've been taught how to do the simple, so you've just got to apply the... this is going to be a harder question for that. So you use your skills from doing the easy part, and then you, you try your best to just get anywhere near to the answer, do the steps individually, and try and just get somewhere close to where you think the answer could be.

AF: Mmm, yeah.

R31M: This is [student R31M]. I think, when it comes to questions being too difficult, I don't think, like [student R36M] said, if you've been taught a part of a topic, I don't think you'll get to a point where they are too difficult. Erm, but I think that's also the job of the teacher to challenge the students and engage them by giving them difficult questions, but instead of then telling them the answer when they struggle, to just give them a pointer or give them a hint to one part of it, because then the next time round they do it, then they'll know to do that thing that they've been told to do. Erm, so, for instance if you get stuck like half way through a question, you don't know that you need to "solve for x" or something for the next part, the teacher might be, "well, find x" and you do pretty much a different variation of that question when they're say asking you what to do the same but with different numbers, and you go, ah, now I need to find that out. And it can be, it can be quite motivating if you then manage to go and do the question on your own, just with that little bit of help instead of being given the answer.

AF: That's more motivating than the teacher giving you the answer?

R36M: I would – this is [student R36M] again – if like, if you're like part the way through a topic and you've done the first half, and you've like done the easier content, I think it would be good for teachers, like, to give a little, like, a challenge question on the board for questions further on in the topic that they think you could maybe achieve, like, if you really, like, think of it, with by using the easier stuff that you've done already, just to like challenge your brain in thinking, what haven't I done yet and what could link to what I've been doing in this last couple of lessons and how, it could be like really easy, when you've been taught it, but when you've got no idea what to do, you've really got to think about it and I think, if you, if they start doing that more, I think that would be better.

AF: OK. Thank you very much. Anybody got anything else to add on that topic, about how teachers could make things, er, more interesting and understandable?

R51M: Yeah – this is [student R51M] – I personally feel like, with some topics, they're like, more engaging than others, like graphs – we often have to spend a lot more time on the board, like having to explain it a bit more, because people like won't understand, like, whether it's a straight line or it's like a cubic graph or where the curve point is, whereas like, other topics like circle theorem at the minute, we're able to have like just a demonstration and then we, like, get on with it. And like, some students as well, are like, people who do further maths, are able to get on with, like, further maths because they've already had a, like, pre-teaching of it. So I feel like, anyone can do anything as long as, like – the question isn't hard as long as you understand what the question's actually asking, what you need to do.

AF: That's a really interesting point. Thank you.

R31M: Yeah, I agree. This is [student R31M]. I agree with what [student R51M] said. And I think it's because we have like, that, again, the variability of the students who are further on can go and do what they want to do – when it comes to further maths they can go on to do more difficult questions. Because some students are more engaged when it comes to working on the board, some students are more engaged if they can go and do their own work, like in their book, and I think because there is that focus between the two, that we're able to do quite well.

AF: OK? That's lovely. Thank you all very much indeed. I'm going to stop the recording now.

[Length of recording: 33'12"]

Appendix F: Focus Group 2 Transcript

Tuesday 20 December, 2022, 10:30am

Year 11 **Middle Set Maths Group**

Researcher [AF]. Students: R73F, R77M, R87M, R96F and R97M

[Introduction and explanation given, including purpose of the focus group and privacy notice.]

AF: So, I'm interested – when you look at an examination question in maths, what d'you look at first, what's your focus? Erm, and what happens after that? Umm, and do you always use the same approach? So those are the sorts of things I'm, I'm interested in. Would anybody like to start me off on a sort of general, a general one? When you look at a maths question, what do you look at first?

R73F: When I look at these ones...

AF: ...This is [student R73F] speaking...

R73F: ...When I look at these ones I like, think that they're like real-life scenarios, so it helps to, like, make them understand more.

AF: Tell me more about that?

R73F: Because, like, when they've got, like, money in them, I find them a lot easier, because they're more like real-life scenarios.

AF: Thank you. Thank you. Anybody like to add to that? Or say something different, about what you look for, what your focus is?

R77M: Er, this is [student R77M] speaking. And the first thing I'd focus on in a maths question is, like, what kind of topics it's bringing in, because I can distinguish what I actually need to do to get, not the entire answer but at least some marks to, for working out – if I can get all the marks, that's great, but that's not necessarily that important.

AF: Right. So why is the topic important?

R77M: Because each, er, each topic has different has different ways to go about things, and if you can work out what topic it is, you can work out the general formation of how to get the answer.

AF: OK. Thank you. Anybody else want to... add their perspective? [Student R96F], what do you think? What do you look for in a question?

R96F: Erm, I think first of all I look for what it's asking us to do. And then I look at what I've actually been given in the question, like what numbers I've been given. Then I just go from there really.

AF: So you're looking for what it asks you to do and what numbers you've been given, and then...

R96F: ...I dinnaa... I just see if I can try different ways of what could work.

AF: Uhah.

R96F: And if I end up getting something where I'm like, I'm thinking I'm along the right track, I just keep going in that direction.

AF: Thank you. OK, now let's have a look at question number 1. So, I'm interested: can you talk me through how you would solve question number 1? [Student R87M], can you talk me through your approach there?

R87M: You find the difference between the two numbers and then you would multiply them by 0.85, then add 150 on top of that.

AF: OK. Now, you said some interesting things there. Can you tell me a bit more about that. Why 0.85, for example?

R87M: 85p is less than a pound, so it would be 0.85.

AF: OK. Right, so, so, umm, your approach there was to subtract the two numbers, yep?

R87M: Yep.

AF: How did you know to do that?

R87M: Er, cos you need to find, like, the difference, because he started off at 27 thousand, then he returns at 28 thousand – so you need to find how many miles he actually drove.

AF: OK. Right. Fine. Now [student R77M], I can see you having a go at that sum already there.

R77M: I'm just getting the general...

AF: ...No, that's fine: you talk... so talk me through what you're, what you're doing there.

R77M: Err, well, what I first did was find the difference between the different mileages – which was what [student R87M] said – and then I times'd it by 0.85 to find out what's the extra mileage he had, and then added a hundred and fifty pounds to that.

AF: OK, and you're confident that, that approach will give you the answer there?

R77M: Yeah.

AF: Yeah. OK. Anybody would do anything different, or would you all do that, that same sort of approach there?

R73F: I'd do the same.

AF: You'd do the same, [student R73F]. OK, thank you. [Student R96F]'s nodding as well.

R96F: Yeah.

AF: What about you, [student R97M]?

R97M: I'd do the same.

AF: Yeh. OK. Thank you. Right. OK. So, now. Would you always use that same approach? [Student R77M], you wrote yourself some notes there just to remind you, kind of, of what to do, the method. Would that, would that approach be the same if you, if you found that in an exam, as if you found it in class, or would you, do you think your approach would be different?

R73F: Erm, [student R73F]. I think mine would be the same.

AF: Good... be the same. Anybody do anything different?

R77M: Errrr... [Student R77M]. My approach would be slightly different, because in lessons I have access to help and I don't have to put so much, like, effort and thought into every question, cos I can ask for little bits of hints, so it doesn't, like, put strain on my brain or anything (I know that sounds horrific). But in an exam, you don't have access to that help, so you'd have to spend more time and effort on the question than you would in lessons.

AF: OK, so in a... in an exam, it takes more time and effort...

R77M: ...Yes...

AF: ...You feel? Does anyone else feel that? [Pause] It's OK, you don't have to agree – he's entitled to his own view there. OK, that's interesting. Thank you. OK, now let's have a look at question number 3. Now question number 1 was on that most people got correct; let's have a look now at question number 3. So, can you talk me through how you would approach question number 3? [Student R96F], can you have a go at that for me?

R96F: Erm, from the way I look at it, I'd approach it as what we call a tree diagram, where we have the percentages of males, and because we want to work out the females who are left-handed, so I'd try and work... obviously it says we have 60% are males, so I'd suggest out of 100%, 40% would be female. If 20% of the males are left-handed, then I'd suggest 80% were right-handed. Erm, and then if it says 21.6% of all people are left-handed, you use the tree to, like, times the, to get the like probability at the end, and from there I'd use that to work out the percentage of the females.

AF: OK. So, [student R96F]'s given us a clear explanation, using a tree diagram. Would anybody else use a different method? [Pause] There are several different methods you could use... you'd all, you'd all use tree diagrams, would you? Yeh. Is that a method you've been taught, [student R97M]?

R97M: Yeah.

AF: OK, so why, why would you think a tree diagram was most suitable here?

R97M: Well, it's like the only way of learning how to, like, work out... [pause]

AF: OK. It's the, the method you've been taught?

R97M: Yeh.

AF: Yeh. OK, erm [clears throat]. Right, now let's have a look at, erm, questions 4 and 5. Now question 4 is, has got more words, and question 5 has very many fewer words, hasn't it? So I'm interested in how your approach differs, and maybe which type of question you prefer, and reasons why – just hearing some, some views about that. [student R87M], would you like to start me off?

R87M: For question 4, I would just do, er, 16,500 times 0.82 to the power zero, cos that'll give you the value when it's new.

AF: Uhuh... OK, talk to me about – this one's got a lot of words; that one's got, just kind of plain numbers, hasn't it, and a few instructions. Erm, d'you find one sort more easy than the other, or harder than the other?

R87M: I feel like when the, when there's less words then I find it easier.

AF: Right? Why's that?

R87M: Cos it's not, like, too much information in a sentence. It's just, like, question 5 is, like, just a, just the numbers, just easier to work out.

AF: OK, now, [student R77M], you were nodding there. What's your view on this?

R77M: Er, I preferred questions more like 5. I don't – I'm just going to be honest here – I do not like real-life problems and stuff like that: it just doesn't sit well with my brain at all. And I, I quite like smaller size questions, because you don't run the risk of not 'getting' what the question's asking you to do, which is pretty much my biggest downfall in maths, because I don't actually read the full question, I just look at the numbers and go, "ok, that's what I need to do." And then at the end there's always something like – I can't give an example – but something that you have to do differently at the end. That's my biggest downfall – smaller words you don't really run that risk, but with stuff like this, you do.

AF: OK. Thank you. What's your, what's your view, [student R73F]?

R73F: I don't mind either: I prefer these ones [points at question 5] but these ones don't bother us [points at question 4].

AF: "These ones" – the ones with just the numbers? But you're happy with the words as well?

R73F: Yeah.

AF: OK. Er, now you were saying earlier on that you quite liked the ones that were in the real-life context. So, as it happens, the real-life one has got the words, hasn't it, and the algebra one has just got kind of numbers and xs and things like that. So you don't have a strong feeling about those?

R73F: No, not really. I just think they're very like simple, short questions.

AF: And does it make any difference if they're got words or numbers?

R73F: Er, sometimes I quite like the ones with words because they're more interesting, so it makes me more likely to do them.

AF: OK. Thank you. [Student R97M], what's your view?

R97M: I'd rather just go for, like, smaller questions – there's less, like, to read.

AF: What d'you mean by smaller questions? D'you mean the marks they've got, or the amount of words?

R97M: Less words. Cos it's, like, quicker to read and you're straight on with the question.

AF: OK, right, erm. [Student R87M], d'you have a view on that?

R87M: I agree with [student R97M]. They're easier to read and easier to get the answer.

R73F: Can I say something else?

AF: Yes, of course, [student R73F].

R73F: When there's, like, a long-winded algebra questions with a lot of marks and no context, I think they're quite hard.

AF: Go on, tell me why.

R73F: There's just a lot more to do and it's, like, it's down to you, because there's not much help with the questions.

AF: Right, OK. So, it's "down to you": the question doesn't help you much. So, um... at that point, do you rely on what you've learnt in a lesson, a method maybe you've learned?

R73F: Cos when you know what you're doing with like these ones it's fine, but when you don't it's not as easy, but with these ones you kind of get help from the words in the questions.

AF: OK. [Student R96F], what do you think?

R96F: Er, I think I do prefer the ones with just numbers, just cos sometimes the words, I can get quite confused with, like, which one's asking us what sometimes. I don't mind the words, but if it was just like the shorter number questions, I think I prefer to do them, cos I feel like I understand them more and I'll just easily get, if I quickly do them then I'll be able to spend more time on the 'wordy' questions.

AF: Cos you... Now that implies that you think that either the 'wordy' questions are harder or that you'll find them harder...

R96F: ...Yeah...

AF: ...Which is it, d'you think?

R96F: I think – with the 'wordy' questions, I think it just depends on the question. Some of them are... If it's like normal little 1-markers then I think I'm ok, but normally when I...

like... where the 'wordy' questions start, like, being 4-markers, 6-markers, then it can get, like, you need to spend more time on them.

AF: Mm-huh. OK, thank you. Anybody else got anything to add on that? [pause] OK, thank you. Now I'm interested, um, in how you think students like you improve your skill in mathematics and, er, I wonder if you can talk to me about this, erm – how did you, how did you develop your mathematical skills, and what did you do to understand better when you came across stuff that was difficult? Anybody like to start me off about how you developed your maths skills?

R73F: [student R73F]. When I, like, first, when we first do a topic, I like normally to go quite slow to, like, process the steps, and then, like, gradually sort of go on and get better.

AF: OK. What do you mean by "process the steps"? That's interesting.

R73F: Because, like, when it's a new topic it doesn't like sometimes click in my head, so, like, I try to put it into steps, so it's quite easier.

AF: And is that something the teacher would do for you, or is that something you do yourself?

R73F: The teacher would do it sometimes... well, the majority of the time [laughs].

AF: And then you'd... er, you'd pick up... how to do it from that? OK, right. Erm. What about you, [student R96F]? How do you think you develop your maths skills?

R96F: Erm, I think, it does depend on the person – if they're willing to work hard and try and understand it. If I didn't understand a topic, I think, 1) I'd ask the teacher – if, like, is there anything extra he needed to go through with us, or I'd just do extra revision material on a certain topic if I didn't understand it, just so, like, it would give us a little bit more support in lessons, so I know just what we're doing.

AF: So you're talking there about practising questions...

R96F: ...Yeah...

AF: ...Lots of times. Is that the way you, you think you best develop your skills?

R96F: Yeah. I think if I just keep practising it, then it'll just sink in. And it tends to be working, so sometimes I think it's ok.

AF: OK. Right. What are, what are the views of the, of the boys on this? [Student R77M], what d'you think?

R77M: Errr, well. I don't – I wouldn't say that I spent a lot of time like developing my skills. I... from... not from a young age, from about Year 6, that's when I really started to get good at maths, just naturally: I didn't do any extra work or anything. Er... I don't do, I don't do any extra work now. But, like I said earlier, when you're in a lesson and you have a teacher who can, like, teach you, teach you physically how to do something... our teacher like gives us different forms of questions, which gives, which gives

different, like – what’s the word? – like, scenarios and, er, the more of them that you can learn how to do like the easier it becomes.

AF: By, by doing different types of questions, that sort of thing?

R77M: By different types of questions on the same topic, yeah.

AF: OK. Thank you. What about you, [student R87M], how did you, how did you develop your maths skills?

R87M: Erm, similar to [student R77M], just the... different, different kinds of formats – I’ve just learnt, learn that way. [pause]

AF: OK. Is that the same for you, [student R97M]?

R97M: Err... I do, like, similar to [student R73F] and [? unclear], I go slower at the start and like, I’d ask the teacher if I need, like, help, and put notes at the, like, front or the back of my book so, like, if I’ve got a, a new equation I’d put, like, it at the front or the back, so I can just, like, flick to it if I need it. And then, like, it’s always there.

AF: OK, so it’s there as a, as a reference for you. Right. So how do you, how do you learn equations like that? Or, or do you not need to because they’re in your book?

R97M: Like, it’d be in my book, and over time if I, like, keep using them, then it’d get, like, stuck in my head.

AF: [laughs] That’s a good phrase – “get stuck in your head”, by keep using it. Does that make sense to other people as well? OK. Nodding all round the table. OK [clears throat] and, finally, I wonder, what d’you think teachers can do to make maths more interesting, more understandable for students? Who wants to start me off on that? [Student R77M]?

R77M: Err... I don’t really know how to make it like more interesting. I find maths in itself just interesting and quite fun to do. So I don’t know, er, how other people think. And so I’m probably not the best person to ask about how to make it interesting.

AF: OK, cos you find it interesting anyway?

R77M: Yes.

AF: All right. [Student R73F]?

R73F: I like maths as well, like, maths is one of my favourite subjects, because I’ve always, like, quite liked it a lot. But um, when we do, like, challenging questions I quite like those, like the long ones, cos there’s, like, a lot of like things you can apply it to, so it’s, like, putting all your knowledge together.

AF: OK, so you quite like it when it’s hard, when it’s more difficult?

R73F: Yeah...

AF: ...Yeah... Why d’you like that?

R73F: I dunno, I just like always being challenged by it. Like, it's just something different [they laugh].

AF: OK. [Student R87M], what's, what's your view? Is there anything teachers can do to make maths more interesting and, erm, more easily learnable?

R87M: No, not really – it's interesting the way it is already.

AF: OK. [Student R96F], what's your, what's your thought?

R96F: I think it's good how they, like, incorporate different types of questions so it's not all, like, just small questions – you've got, like, the bigger questions and they're not, like, maybe all like the different types of questions. Like, if it's a certain topic, they'll, like, change it around a bit, the way the question's formed or something like that... It just makes it different, so you're not doing the same repetitive thing each lesson.

AF: And that variety...

R96F: ...Yeah...

AF: ...You think, helps? OK, thank you. [Student R97M], what d'you think teachers can do to make maths interesting and understandable?

R97M: Err, [unclear...] understandable, like how the teachers, like, go through all the questions. You can't really make a lesson, like, fun... it's always going to get, like, boring at some points, and more fun at other points.

AF: And that's just how it is?

R97M: Yeh.

AF: OK, so nothing teachers can do about that, particularly? OK. Thank you. Is there anything anybody else wants to say about maths or maths questions or things that make, that they find difficult, or things that make questions easier?

R77M: Er, I have actually thought of something.

AF: OK, [student R77M]...

R77M: So I am, I am going to go back to what [student R73F] said, and the more challenging questions are the most interesting and, er, I don't know how to word this, but when they're really, really hard and not many people get it, there's a select few of people who will, like, get it, like, the first try. And I, er, I prefer it when it's like that – there's then, like, I have the ability to, like, help those around us, [student R87M], for instance: [student R87M] quite often asks us for help. I enjoy that – I enjoy actually helping people: that's fun to me.

AF: OK, so you like it when it's actually harder and maybe you're better at it than some other people.

R77M: It sounds egotistical when you put it that way...

- AF: ...Well, not really, but you, like [student R73F], maybe you enjoy the challenge of it but then you're able to explain it to other people.
- R73F: I think sometimes it does help to, like, explain things to other people, like, cos it, like, it shows that you understand and it's like telling yourself you understand.
- AF: So, explaining it to somebody else – does that help you to understand it better?
- R73F: Sometimes. Like, if we're going through a question and I sit next to my friend and we're going through a question it, like, helps me understand if we're going through it, like, together...
- AF: ...Ahah...
- R73F: ...Cos it shows we both know something about it.
- AF: That's really interesting. Thank you.
- R73F: To get another person's perspective's good as well.
- AF: Mm-umm...
- R73F: ...Cos I think differently to how my friend thinks – like, everyone thinks differently, so when you look at it from a, like, different point of view, that sometimes helps you.
- AF: OK. Thank you very much. Thank you. Anybody else got anything to add? That's been really interesting. Thank you so much for your contributions: I shall look forward to listening to those again as I transcribe them. [End of recording.]

[Length of recording: 21'08"]

End of thesis.