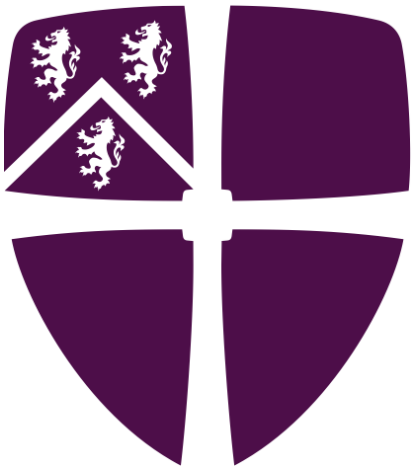
MEASURING CHANGE IN MALTESE EDUCATION: AN IMPACT EVALUATION OF LARGE-SCALE POLICY INTRODUCTION ON LEARNING OUTCOMES AND THE QUALITY OF EDUCATION.

DOUBLESIN, GLENN

Measuring change in Maltese education: An impact evaluation of large-scale policy introduction on learning outcomes and the quality of education.



Glenn Doublesin

School of Education
University of Durham

This dissertation is submitted for the degree of Doctor of Philosophy
June 2023

Declaration

This thesis is the sole work of the author and does not include any material resulting from other person's work or through collaboration with others. No material contained in the thesis has previously been submitted for a degree, diploma or qualification in Durham University or any other institution.

Abstract

This research investigated the impact on the quality of education across the Maltese Islands of two broad-scale educational policy changes — the National Minimum Curriculum (1999) and For All Children to Succeed (2004). It did so by investigating change in student outcomes on a set of five benchmark tests held across the same period as the policy introductions.

The initial stages considered a quality education framework (*input – process – output – context)* to underpin the key processes of the study and help define the educational functions being investigated. The main part of the analysis then established a timescale based on the NMC and FACTS polices before analysing changes in examination constructs and forms. Once the inputs, processes and contexts were structured, the research proceeded to investigate outputs using Junior Lyceum Entrance Examinations results data. The overlaying of time series analysis for each of these analytical processes from before and after the policy introductions allowed the determination of impact points and direction of effect.

The research indicates that the general effect of the two policies had an indirect positive influence on student outcomes. More specifically, there was/were:

i. an overall rate of improvement in student achievement scores before and after the introduction of the NMC and FACTS policies.
ii. discrepancies between application of conjunctive criteria (used by the EAU) and a parallel analysis of the data using compensatory criteria suggest that the criteria applied to the Junior Lyceum Examinations would have skewed signs of progress in teaching and learning at classroom level.
iii. an increase in the rate of change of achievement when comparing the decade before and after the introduction of the NMC. The main changes took place at the beginning and end of the decade with the period 2001 – 2008, showing very slight change when using the compensatory data sets.
iv. English, and mathematics, had a disproportionately higher impact on student success rates that made the inclusion of the other three exams practically irrelevant.
v. The Junior Lyceum Examination set as a whole lacked independent objectivity and as such could not be considered an effective benchmarking tool.

The aim of the study was to evaluate the effectiveness of the reforms on student learning, develop tools for monitoring change outcomes and identifying key factors driving the change. By analysing changes in student achievement and identifying the contributing factors, the study also sought to provide insights for future policy decisions and educational reforms.

Acknowledgements

I would like to acknowledge several important individuals without whose support this work and my personal growth in this field would not have been possible.

Mr Louis Scerri and his team at the Educational Assessment Unit for their interest and support in this research and their assistance in locating and scanning the volumes of records required for this research.

Prof Mark Borg for inspiring this work and his confidence and support as the research progressed.

Professor Stephen Gorard for always getting straight to the point and grounding me in the realities of academic research and scrutiny.

Professor Steve Higgins whose guidance allowed me the liberty to discover my path through this research and the confidence to follow it.

Finally, Caroline, my indisputable better half, for her continuous patience, care, and encouragement throughout the lengthy process of putting this thesis together.

The research work disclosed in this publication is partially funded by the ENDEAVOUR Scholarships Scheme funded through national funds.

Contents

List of Tables

List of Figures

List of Abbreviations and Acronyms

Acronym	Name
ARI	Automated Readability Index
CID	Cognitive item Demand
CL	Cognitive Load
CLI	The Coleman-Liau Index
CLT	Cognitive Load Theory
D_i	Discrimination Index
DQSE	Directorate for Quality and Standards in Education
DQSE	Directorate for Quality and Standards in Education
EAU	Educational Assessment Unit
EER	Educational Effectiveness Research
EFA	Educational For All
EIPP	Evidence-Informed Policy and Practice
EN	English (Subject)
F	Facility Index
FACTS	For All Children to Succeed
JLEE	Junior Lyceum Entrance Examinations
KPI	Key Performance Indicator
LIX	Lasbarhetsindex
LOS	Learning Outcome Statements
MCQ	Multiple Choice Question
MEDE	Ministry of Education and Employment
MEYE	Ministry of Education, Youth and Employment
ML	Maltese (Subject)
MS	Microsoft
MT	Mathematics (Subject)
NCF	National Curriculum Framework
NMC	National Minimum Curriculum
OCR	Optical Character Recognition
OECD	Organisation for Economic Co-operation and Development
PDF	Portable Document Format

PISA	Programme for International Student Assessment
QI	Quality Indicator
RL	Religion (Subject)
RQ	Research Question
SEC	Secondary Education Certificate
SS	Social Studies (Subject)
TIMMS	Trends in International Mathematics and Science Study
UNESCO	The United Nations Educational, Scientific and Cultural Organization
XLXS	Microsoft Excel Open XML Format Spreadsheet file
Δ%	(%pass - %fail)

1 Introduction and context

This research investigates the impact of large-scale educational policy changes on the quality of education across the Maltese Islands. More specifically, it aims to explore whether it is feasible to measure the effect these changes may have had on students' learning outcomes through a longitudinal analysis of the Junior Lyceum entrance examination results and associated reports from 1997 to 2010. This examination set spans two major national policy introductions and will allow the study to determine if there were, in fact, any subsequent detectable effects following the introduction of these policies.

The research will draw on a comparison of change in achievement and attainment, and the magnitude of those changes, to consider associated implications for the quality of education.

1.1 National policy documents being considered

Over the last two decades, Malta's education system has undergone some far-reaching policy changes intended to have a positive impact on a wide range of educational areas (Attard Tonna & Bugeja, 2016; Borg & Giordmaina, 2012; Cassar, 2021; Galea, 2004; D. Mifsud, 2015; Mizzi, 1999). Each of these changes integrated particular strategies to raise the quality of education across the islands and explicitly stated this as one of the principal objectives of the new policies (MEDE, 2012; MEYE, 1999, 2004a).

The policy documents establishing these changes each defined a historic developmental stage in the Maltese educational system to make systemic changes to pedagogies and concepts of teaching and learning (Attard Tonna & Bugeja, 2016; Cutajar, 2007; Galea, 1999, 2004). Grima et al. (2008) stated that they also intended to institute certain paradigm shifts in the outlooks, attitudes and practices of the nation's education profession, although the pace at which such

changes were implemented would prove more challenging than initially expected (2008, p. 29).

Policy-driven reforms continued with the rollout of the National Curriculum Framework (NCF), which has been implemented over the last few years, setting the framework to initiate outcomes-based teaching and assessment starting during the scholastic year (2018 – 19).

The principal documents driving these policy changes over the last two decades for primary and secondary education have been:

- The National Minimum Curriculum: "Creating the future together, National Minimum Curriculum" (NMC) (Ministry of Education, Youth and Employment) (MEYE, 1999)
- The setting up of School Networks to facilitate the decentralization of education across the islands and the transition from the primary to the secondary cycle: "For All Children to Succeed, A New Network Organisation for Quality Education in Malta" (FACTS) (MEYE, 2004a)
- A National Curriculum Framework (NCF) establishing the use of learning outcomes as a basis for education: "A National Curriculum Framework for All" (Ministry of Education and Employment) (MEDE, 2012).

Although the NCF does not fall within the scope of the research being undertaken here (1997-2010), it still stands as a follow-on to national policy change on the same scale as its predecessors and contrasting analysis could shed light on how the NMC and FACTS are linked to the NCF and how they may have influenced or led on to the NCF and its intended outcomes.

1.2 Structure of the Maltese education system

The educational sector in Malta comprises three main sectors: the State, the Church, and the independent sector. The State sector represents the largest portion of the educational system across the islands with state primary schools that were co-educational, while State secondary schools were single-sex schools. Within this sector, all schools operated as non-continuous schools and students needed to transition from primary to secondary education in different schools (Grima et al., 2008; MEYE, 1999, 2004a).

The state secondary school system was established to run along two streams: Junior Lyceums and Area Secondary schools. Admission to Junior Lyceums was based on the results of the Junior Lyceum entrance examination, and students are allocated to the Junior Lyceum located within their catchment area. On the other hand, students who either do not sit for the JL

examination or do not pass it are admitted to the Area Secondary school in their respective area. These catchment areas became Colleges after the FACTS policy was implemented.

The Church sector comprises schools that are run by religious organizations and offer education based on Catholic principles. These schools can be found throughout Malta and Gozo, providing an alternative to the State sector. Some of these schools are continuous with students progressing from primary to secondary, whereas other church primaries were not, with a number of their pupils sitting for the JLEE.

During the period in questions, the independent sector included private schools that operated independently from both the State and Church sectors. These schools often had their own specific curriculum and admission criteria and many, but not all, offered continuation from primary to secondary schooling. Some of these independent schools did not, however, offer secondary schooling and a number of their students used to sit for the JLEEs as one of their options to transition into secondary schooling.

During the 1997 academic year in Malta, compulsory education (pre-16) saw a total enrolment of 64,757 students. This figure was divided between 35,261 students in primary education and 29,496 students in secondary education. Fast forward to 2008, there was a noticeable decrease in the student enrolment in both education levels. The number of students in primary education dropped to 26,772, while those in secondary education declined to 25,793 (*Education Statistics 2006*, 2010).

It is worth noting that the proportional distribution of students between state-run and private schools (including both church and non-church institutions) remained steady over this period, maintaining an approximate ratio of 72% in state schools to 28% in private institutions.

1.3 The lack of quantitative evidence to inform decision making

This study uses the terms achievement and outcomes to refer to two types of outputs resulting from the educational process. For the purpose of this paper, achievement rates are associated with individual achievement on standardised tests, while outcomes are intended to refer to overall attainment in the form of pass-fail rates based on established success criteria (Scheerens et al. 2011a, p. 35). This study investigates variations in both achievement and outcomes, and due to the pass-fail rates being dependent on achievement in five different subject exams there is some overlap in the use of the terms. Both terms, however, constitute the main analytical components of the output domain discussed later as part of the quality framework.

Considering that the introduction of the NMC and FACTS brought about such major changes to the entire structure of education in Malta, there is a limited number of systematically designed empirical research actions that investigate the policies' respective impacts on learning outcomes. The topic has not been sufficiently reviewed or researched as it pertains to quality for all, student achievement, or attainment. Kreber & Brook (2010) point out that this shortcoming in systematic empirical evaluation is not unusual.

> *"Although most educational development professionals value the importance of monitoring their programme's impact, systematic evaluation is not common, and often relies on inference measures such as extent of participation and satisfaction." (Kreber & Brook, 2010, p. 96).*

Similarly, Sebba (2003, p. 16) agrees with Gorard's 2001 assertion that there is a shortage of researchers with the skills and technical proficiency to *"interrogate these databases which are underdeveloped"*. This latter assertion may be specific to the UK context but is a reflection of more global realities that exist in other countries including Malta. Valenzuela, Bellei, & Allende (2016) state that their research into school change highlighted that *"it is a comparatively underdeveloped issue in the literature"* (2016, p. 1).

Qualitative studies of the impact of the NMC and FACTS on teachers and school management personnel are more readily available; Cutajar (2007), Mifsud (2015) and Fenech Adami (2004) to mention a few, and internal surveys and reviews have established broadscale perceptions and general attitudes towards the changes. However, these investigations were unable to state definitively if student learning quality, outputs or outcomes were in fact affected in any way as a result of the introduced policy. One particular study by Grima et al. (2008) makes an association between modernising the syllabus and positively impacting examination design.

White (2010, pp. 152–164) points out that when large scale impact studies are conducted, looking for patterns in the results and analysing the statistical changes — applying a quantitative approach — is both appropriate and more effective in rendering an objective picture of the resulting transformation. Similarly, Bird, Anderson, Anaya, & Moore (2005) argue that a quantitative approach is preferable when identifying effect in student learning, although a mixed approach gives a more holistic understanding. Nevertheless, though such objective measurements of the impact on learning outcomes are recognised by policymakers and decision makers associated with the NMC and FACTS (MEYE, 2004a, p. 44; Mizzi, 1999), it is not reflected in any related publications or research. Monitoring systems, which inform the policymakers as to whether or not there is convergence towards or divergence from the anticipated outcomes, are also lacking. Investigation of the research shows that when impact

analysis research is conducted, it is mainly focused on targeting very specific criteria. Said (2015) for instance, investigates the influence of teachers' behaviour on mathematics outcomes for six-year-olds.

1.4 The need for evidence-based decision making

Policy monitoring and evaluation are applied to varying degrees of effectiveness, but most are concerned with assessing impact (Stevenson, 2003, p. 36). Weimer & Vining (2017) highlighted the importance of integrating robust policy evaluation mechanisms during the policy design and development phases as a key element to ensuring proper implementation and accountability by those doing the implementation. The OECD Review on Evaluation and Assessment Frameworks for Improving School Outcomes - (OECD, n.d.) states that such mechanisms are necessary in order to support effective system evaluation.

The requirement to monitor and gauge outcomes and progress is in fact stipulated by the policy documents in question, although these requirements are structured to target individual student achievement and deemed to be the responsibility of the schools or colleges.

> *"Results will help recognise better the students' achievement and the support needed for those who do not reach the expected targets. This will lead to further inclusion of students in the same school. It is important to note that in any educational path, it is imperative to stop and take stock of progress and what next steps are needed." (MEYE, 2004a, p. 44)*

There were, however, no systemic structures put in place that would monitor the overall change in outputs or outcomes, nor any suggested mechanism to do so in order to understand the impact of these new policies on student learning.

In contrast to the NMC and FACTS policies the latest national educational policy being introduced, the NCF, specifies a review of targets of achievement (MEDE, 2012, p. 24) to enable data-driven decision making and states this by explicitly stipulating that it is necessary to assess the impact of the changes on attainment.

> *".... regular reviewed targets of achievement that will enable education leaders and policy makers in education to assess the impact of the NCF on the attainment of the Secondary Education Certificate at MQF Levels 2 and 3 (SEC) and the Secondary School Certificate and Profile at MQF Levels 1 and 2." NCF (MEDE, 2012)*

The implication here is that an effective understanding of what is happening due to introduced policy is only seen in terms of outcomes (students completing secondary school; qualification statistics; literacy rates) and as a result can only be observed over a longer term. Any

immediate and unintended influences generated by the introduction of the policies will only be realized at a later stage, making it difficult to adjust for unintended consequences (Baker, 2013, p. 85; Caruana & Allied Newspapers Ltd, 2016).

Considering that the newly designed learning outcomes (*DQSE : Directorate for Quality and Standards in Education*, 2016) are subject and age/grade level specific, their design is intended to allow better-targeted improvement in learning outcomes (MEDE, 2012, p. xiii). With these learning outcomes underpinning a new direction in Maltese education, it should therefore be possible to measure the impact on student learning on a micro-level according to subject, age group, demographic, and by school or college and with more immediate results. If raising the quality of education for all is a tenet repeated with every major policy change, then measuring learning outcomes in terms of attainment alone is neither quick enough nor comprehensive enough to inform the implementation process.

1.5 System performance and available data

Ideally, a policy impact analysis would cover a broad range of high quality structured data sets at different levels within the system that can be applied to draw inferences and determine system performance (*OECD Review on Evaluation and Assessment Frameworks for Improving School Outcomes - OECD*, n.d., p. 8). The OECD review further states, however, that successfully aggregating a broad range of "measures of system performance" is a challenge in most situations and the availability of the data, or lack thereof, should not be the sole determining factor (p. 8). The review goes on to state that student assessment is a key element in assessing system performance as it provides evidence of actual student performance compared to projected or desired performance.

In the context of this research, the possibility of analysing outcomes data can only be done effectively through an analysis of attainment statistics for the yearly Junior Lyceum entrance examination supplied by the Educational Assessment Unit (EAU). Although public school examinations are centralised, accessing these summative in-school results has proven difficult to accomplish as there is no centralised system for the collection and collation of these results. Such records are kept by the schools at their own discretion and according to their own individual procedural structures. As a result, these data sets are not likely to be reliably consistent or appropriately representative (Scott, 1990).

The EAU record-keeping procedures for the Junior Lyceum Examinations have, on the other hand, been consistent throughout the years and a complete record has been kept dating back to the late 1980s. Furthermore, it was understood that although not all the Year 6 students

across Malta and Gozo chose to sit for these examinations, the majority of them did actually do so, establishing an increased level of reliability in the analysis and allowing stronger inferences to be drawn from their interpretations (Scott, 1990). The Department of Curriculum Development, Implementation and Review & Educational Assessment Unit (2005, p. v) recognised these examinations as "an important benchmark in our educational system".

The consistency of the EAU records and their repetitive structures over the years allows them to underpin comparative longitudinal analysis and gather insights through the respective statistical variations from year to year.

1.6 A more definitive impact assessment

The Maltese context for understanding the impact of far-reaching national policy changes on student learning has been recognised by the policymakers and stated in the policy documents (MEYE, 1999, 2004a). Despite these stated intentions however, a measured evaluation of the effect of these policies has been predominantly conducted through qualitative means (Borg & Giordmaina, 2012; Grima et al. 2008), with limited analysis of an *"impact on candidates"* (Grima et al. 2008, p. 69).

Questions remain though: Was there a quantitatively measurable effect on candidates learning outcomes? Within the Maltese context, can any such impacts be measured objectively? Considering that this would be a retroactive study, can any effects be quantified, or effect sizes determined, due to the complexity of large-scale policy introductions?

Although associated studies, research and reviews have been conducted over the policy implementation periods, none of them have offered a quantifiable comparison of system performance before and after the policy introductions. A more definitive and detailed view of policy impact can be better rendered through a combination of qualitative and quantitative assessment methods (Bird et al. 2005, p. 359) and it is this latter evaluative method that is needed to expand our understanding in this context.

1.7 Thesis structure

The study is arranged in four sections that help organise the key chapters and general transition of the thesis. The first section presents a literature review over two chapters developed to present key concepts that underpin the subsequent methods and analysis sections before the closing section presents the overall discussion and conclusions associated with the thesis.

Part one of the literature section discusses concepts associated with evidence-informed policy and practice (EIPP) and impact-monitoring systems that are used for large-scale policy implementation. Furthermore, it discusses a quality framework for analysis (input – process – output – context) that structures the subsequent discussions and impact evaluation, linking policy to implementation processes and to any variation in outcomes.

Considering that the results and outcomes of this research cover a span of fourteen years, the second part of section 1 is a review of how examination standards tend to vary over time. This is done from the aspect of test constructs and forms and reviews the literature of how their nature might vary longitudinally to affect difficulty of what should be parallel examinations.

Section 2 presents the research design and methodologies over three chapters. The complexity of policy impact on examination outcomes has led to an intricate web of analytical methodologies, each considering a particular aspect of outcomes or achievement and each structured on a longitudinally comparative framework. The first of these, chapter 4, reviews the methods required for documentary and data analysis and links them to the research questions that underpin this study. It also lays out the sequence of analysis and maps the overall analytical process. Chapter 5 focuses on the longitudinal and cross-sectional structures developed and applied as a critical analysis of year-on-year contexts, achievements, and outcomes. The last chapter of section 2 offers a detailed description of the digitisation processes applied to support the respective analysis. The actions described in this last chapter were essential, considering most of the data and information was only accessible as printed hard copies rather than digital soft copies. The accuracy of digitising this information was key to maintaining the reliability of any analysis that followed.

The analysis and results are presented in section 3 over three chapters. This section is informed by the research questions and guided by the quality framework for analysis proposed in the literature review. Chapter 7 considers inputs and context, and is a review and analysis of the discourse used in the two key policies (National Minimum Curriculum (MEYE, 1999) and For All Children to Succeed (MEYE, 2004a)). As such, it describes the input and contextual scenarios influencing the educational landscape at the time. Chapter 8 looks at the process and context associated with the Junior Lyceum Examinations and presents a trend analysis of various exam characteristics: construct continuity, test form complexity, and consistency in difficulty levels. The longitudinal analysis of these trend variations was designed to be compared to similar longitudinal analysis of outputs over the same period in the form of student achievement and outcomes. This latter set of analysis are presented in the closing

chapter of this section, chapter 9. Combined together, the different sets of time-series analyses from these three chapters offer a comparative facility that can link any variation in student outcomes to possible affecting factors in other areas of the study.

The concluding section of the thesis will offer a discussion of the outcomes of the analysis and results in the context of the research questions and relate conclusions to the literature. The chapter concludes by considering the limitations of this research, and also notes the importance of instituting systems that monitor progress in achievement to reflect on policy impact on the quality of education.

Section 1: Literature Review

Section Overview

The literature review is organised into seven sections discussed over two chapters and intended to present key concepts associated with effective large-scale policy processes in education while laying the groundwork for a framework on which to assess the impact of such processes. This review will support the subsequent development of a framework for analysis to evaluate the impact of the two large-scale policies introduced to the Maltese educational landscape in 2000 and 2005.

The first chapter in this section reviews concepts associated with evidence-informed policy and practice (EIPP) and systems for monitoring and evaluating large-scale policy impact. The second chapter then focuses on the longitudinal sustainability of examination standards for comparative purposes. Furthermore, it reviews concepts associated with test constructs and forms with a more detailed consideration of how inherent characteristics associated with such constructs might fluctuate to affect varying difficulty levels on parallel tests forms.

The two parts of the literature review will underpin the general processes and principles applied in this study namely:

i. a longitudinal analysis of outcomes from an annual standardised test compared to a timeline analysis of policy introduction and implementation.

and

ii. a deeper longitudinal analysis of continuity and consistency of test constructs and forms together with an understanding of variations in mental load and difficulty levels that may also have had an impact on the overall student outcomes.

2 Educational effectiveness and quality.

2.1 Chapter overview

The initial section of the first chapter is intended to establish a general description of what is considered to be an ideal state of affairs associated with broad-scale EIPP in education. The chapter begins by exploring literature associated with EIPP specifically regarding large-scale policy processes, factors that impact the associated decision-making practices, and how such policies are best applied to improve educational systems. This first section will work to inform the interpretation of any realities and relationships associated with the research into the effectiveness of the large-scale policy changes that constitute the basis of this study.

The second section focuses on literature discussing educational effectiveness research intended to deliver insights into systems for measuring educational improvement and their application in informing the decision-making process. This section is intended to understand systems applied to sustain effective reform over prolonged periods and how those systems best support the implementation mechanisms linking large-scale policy development at the governmental level, to school and classroom action. More specifically, this section reviews data-driven information and monitoring systems as these are considered key types of support structures that work to inform ongoing decision-making processes. These systems use types of data to inform such support structures — particularly longitudinal data — that make them particularly relevant to this research.

The last section of this chapter establishes a framework to underpin a structured evaluation of policy impact on outcomes, in particular on student achievement. The framework is intended to reinforce the analysis and link introduced policy to the process of implementation and consequently, the resulting outcomes. Establishing such a linkage through an appraisal of the

associated literature and how it has been applied in international institutions, will allow the analysis to rest on a framework that will underscore arguments of association between changes brought about by policy at the input stage and effects seen at the output. This is not however intended to ascertain the arguments being made but to add support to the association from input, to process, to output.

2.2 Evidence Informed Policy and Practice (EIPP)

This section outlines the recognised importance of EIPP and its links to maintaining standards and quality in educational systems. It highlights how most modern educational systems have integrated EIPP and connects this research to a recognised and ongoing need for continuous monitoring of outcomes to inform policy development and implementation.

2.2.1 Background

The literature points to an increasing recognition of EIPP as an effective tool in improving the quality of education. A 2013 OECD review of evaluation and assessment frameworks found that governments and education policy makers were making more regular use of both to monitor and improve outcomes in education. The report associated this increased use of evaluation and assessment with a number of different emergent factors including a more defined set of requirements for "*effectiveness, equity and quality in education*" and a greater need to make use of *"evaluation results for evidence-based decision making"* (OECD, 2013, p.13). There has also been a sustained push across the European Union to support evidence-based policy making to improve education standards across the continent (European Commission. Education, 2017).

Adams (1993) and Barber (2010) have argued that the emphasis on "quality" education as an important part of a modern growing society has been receiving more attention and importance in the more developed economies since the 1970s and '80s. These efforts have persisted and are seen as an ongoing commitment to quality education (Boeren, 2019) by governments and international institutions that make use of quality statements to underpin performance indicators for all aspects of the educational system (OECD, 2013; UNESCO, 2000, 2002, 2005, 2017). Similar emphasis is reflected in the Maltese context as stipulated in the national education policy documents from the last two decades (MEDE, 2012; MEYE, 1999, 2004a).

There has been a growing awareness and recognition of the need to inform policy and decision-making processes using EIPP. Higgins (2020), Pellegrini & Vivanet (2021) and Slavin

(2020) have noted the recent rapid growth in evidence-based education policy and reform, with Lewis & Hogan (2019) and Pellegrini & Vivanet (2021) remarking that policies are being designed to integrate and support the use of such evidence in the decision-making process.

Although controlled experiments are considered a strong primary source for such evidence-based decision making (Slavin, 2020), evaluation and assessment frameworks have also been employed as key sources of evidence that establish, maintain and continue to develop quality educational systems (Iwu et al. 2018; White, 2020). Such evidence is employed not merely as an outcomes indicator, but as an integrated monitoring system that allows all stakeholders to take stock of past, immediate, ongoing, or projected quality outcomes of education.

> *"In all countries, there is widespread recognition that evaluation and assessment frameworks are key to building stronger and fairer school systems. Countries also emphasise the importance of seeing evaluation and assessment not as ends in themselves, but instead as important tools for achieving improved student outcomes." (OECD, 2013, p.20)*

2.2.2 Politicked research

Although EIPP is an accepted modus operandi influencing a broad swathe of government policies around the world, its impact (especially that resulting from international assessment outcomes) tends to skew a truly objective decision-making process relevant to national or regional contexts or requirements (Lewis & Hogan, 2019). Nonetheless, application of more relevant and targeted EIPP can make valid and significant contributions if underpinned by relevant purpose.

Progressive education systems today have grown ever more dependent on tangible evidence to inform decision making within every related sector and across each branch within those sectors (Anderson et al. 2003; Bush, 2002; Cooper et al. 2009; Hamersley, 2008; OECD, 2013; Stone, 2001; Whitney & McIntosh, 2001). However, Cooper et al. (2009, pp. 160–161) and Avis et al. (2014, p.46) highlight concerns in the discourse that criticise evidence-based decision making as technocratic, limiting and tending to give a false sense of objectivity. The critical discourse and debate on the matter is not whether or not this data-driven progress is the best way forward, that is an accepted state of affairs. Rather, the discourse tends to focus on "which research?" is being supported; "why?" and "how?" is certain inquiry supported over other sources of empirical evidence; and "who decides?" on which progress is important and needs research support (Cooper et al. 2009; Morrison, 2007, p.13; Desforges, 2003, p.5; Avis, Fisher, & Thompson, 2014).

Similarly, Gorard, See, & Siddiqui (2017) have reservations about the usefulness of educational effectiveness research due to the lack of an underlying "strategic direction" that would serve to allocate better research support towards priority areas. Furthermore, both Gorard et al. (2017), Cooper et al. (2009) and Slavin (2020) point to external sway from funding entities, lobby groups or policy departments that, inadvertently or not, redirect progress-through-research according to their own agendas rather than working with broader purpose.

Nevertheless, putting aside both these discussions about particular issues of politicking and possible agenda-driven selection methods influencing which research or datasets are chosen for consideration — looking at the matter from a purely pragmatic perspective, from a functionality perspective — decision making, policy structuring, planning, implementation and monitoring systems, and gauging of performance are inherently integrated with systems that draw on research and empirical evidence to inform their function and improve any outcomes derived from their use. This is relevant at the school or college level as discussed by Bassey (1999) and Coleman & Lumby's (1999) discussion of the importance of site-based practitioner research, but also on broader institutional scales as pointed out by Cutajar (2007), Pellegrini & Vivanet (2021) and Cooper et al. (2009)

> *"The rationale for the use of evidence is obvious. Using research evidence should lead to more informed policy, higher-quality decisions, more effective practices, and, in turn, improved outcomes." (Cooper et al. 2009, p.160)*

Bassey (1999, p.39) points out that evidence-based research is what leads to *"Critical enquiry aimed at informing educational judgements and decisions in order to improve educational action.",* while Honig & Coburn (2008) emphasise that data-informed systems will help *"ground educational improvement efforts"*. Pellegrini & Vivanet (2021) have also argued the importance of evidence-based policy and practice and that interest in evidence-based education has recently been growing exponentially. This concept of grounding transformational efforts brought about by the introduction of national policies on empirical evidence underpins the educational effectiveness research which is discussed below.

According to Cooper et al. (2009) and Cutajar (2007), empirical evidence is the instrument that informs proactive or reactive measures needed in following up the implementation process (Honig & Coburn, 2008; Sebba, 2003). Furthermore, the last couple of decades have seen a definitive move by governments around the world to try to integrate evidence in decision-making frameworks and a general realisation of the importance of this integration (OECD, 2013, p. 13; Pellegrini & Vivanet, 2021; Whitehurst, 2004). Cooper et al. (2009, p.163) state

that *"Government policy documents in many countries now make explicit mention of the importance of research in formulating policy."* and the Eurydice report (European Commission. Education, 2017) encourages and supports such actions across member states.

Questions about "how decisions are taken", "who decides?" and "to what end?" will remain part of the ongoing discourse in this field. The conditioning of educational research by politicked agendas is, for the most part, unavoidable, as funding is allocated by those responsible for strategic and political decision-making (Slavin, 2020), and therefore by those with an agenda to meet. More pertinent to the context being discussed here, and relevant to a national-level discourse, is the consideration of whether or not applicable information garnered from educational research efforts is subsequently and effectively being applied to future planning and implementation practices through established feedback mechanisms and operational frameworks.

2.2.3 School improvement systems: What is the research indicating?

School improvement systems have, over the last few decades, been steadily shifting from school-based initiatives to more centralised action directed by governmental institutions, with an increased degree of systemic effectiveness but lesser responsiveness to particular school needs.

A review of school and system improvement literature by Hopkins, Stringfield, Harris, Stoll, & Mackay (2014) outlines a gradual shift in school improvement efforts over the last five decades, from individual school initiatives to system-wide reform starting at the national, regional or district level. They go on to say that with large-scale change initiatives, although there is a reduction in specific and contextual considerations associated with individual schools, the probability of progress being made is actually strengthened by systemic interventions. Hopkins et al. (2014) point to a review of research by Nunnery (1998) who noted that although externally implemented change was inconsistent in its outcomes, internally initiated changes at the school level were *"even less likely to result in achieving initially desired outcomes."* (Hopkins et al. 2014, p.263).

An earlier review by Smith & O'Day (1990) refers to research by Clune, White, & Patterson (1989); Fuhrman, Clune, & Elmore (1988); and Mullis & Jenkins (1990) to highlight similar findings in school-based reform initiatives, and notes that in this regard, *"…evaluations of the reforms indicate only minor changes in the typical school, either in the nature of classroom practices or in achievement outcomes."* (Smith & O'Day, 1990, p.233)

Nunnery's 1998 study of educational innovation and reform looked at several initiatives undertaken within the US and the large-scale evaluation studies associated with each of those initiatives: Aiken (1942); Berman & McLaughlin (1974); Bodilly (1996); Crandall & Loucks (1983); and Stringfield (1997). Nunnery's review drew on these investigative evaluations to compare the effectiveness of reform implementation when the reform was developed externally as opposed to being developed in-school. Drawing on the comparative evaluation within each of these studies, Nunnery (1998) noted that locally developed initiatives to introduce and implement changes had a lesser probability of impacting student outcomes than changes that came from external initiatives (Nunnery, 1998, p. 285)

The consequences of these findings have had an impact on how efforts to improve student achievement have been conceived, designed, and implemented. England, Scotland, Canada and Holland have all taken on board key findings from Nunnery's 1998 review when designing such reform interventions. Hopkins et al. (2014, p. 264) note that the means of getting these initiatives to improve student achievement were national policy-support mechanisms in these countries with the expectation that they would be more efficient at affecting change than in-school initiatives.

There are various arguments put forward to explain the reasons for the difference in effectiveness between systemic and in-school improvement systems. Borman, Hewes, Overman, & Brown (2003) have stated that the stronger organisational capacity of larger institutions is what gives them the capability to better structure and implement reform efforts. Furthermore, they argue that larger institutions are better able to maintain support teams and resources that would shore up efforts over the longer term. Another analysis by Healey & DeStefano's (1997) considering reform in schools, put forward arguments similar to Nunnery's (1998) that although schools have the capacity to implement systems that lead to development and improvement within their own institutions, these reforms will face *"implementation challenges associated with sustaining reform…"* (Healey & DeStefano, 1997, p.8). This particular point is drawn from experiences in reforming education in developing countries but is extrapolated to make an argument for similar outcomes in developed countries (and more specifically in the U.S.).

Extending the idea that institutional size matters in sustaining effective reform, Anderson & Holloway (2018) and Pellegrini & Vivanet (2021) state that the shift from localised to national level was not an end in itself. Their work argues that there now seems to be a further transition to a globally-influenced approach in educational policy design and creation. The

reports mentioned earlier, presented by the European Commission. Education (2017) and European Commission/EACEA/Eurydice (2017) reflect such international efforts to nudge national governments towards similar, if not common, evidence-informed policy directions. Similarly, work by UNESCO ((UNESCO, 1990, 2000, 2015, 2017) indicates that efforts to shape national policy globally (with a greater impact on developing nations) has been taking place since the early 1990s. However, although there may be benefits drawn from national-level policy development, work by Lennon (2016) suggests that there are risks that things can go wrong if implementation processes are not properly realised, resulting in no tangible effects.

The key point remains reflected in the discussions presented by Borman et al. (2003), Healey & DeStefano (1997) and Nunnery (1998), that sustainability in reform movements can be better served through centralised (albeit bureaucratic) systems of needs analysis, planning, implementation and support structures. This shift from the *school as the unit of change* (Hopkins et al. 2014; Smith & O'Day, 1990) — ascribed to in the late 60s, 70s and 80s — to an educational development model that is more centrally-driven and targeting multiple-school communities, places the onus of change and educational improvement on central governing bodies (Datnow & Park, 2010; M. S. Smith & O'Day, 1990). Implementation processes are directed through education systems that govern a multiplicity of schools and could be, perhaps tend to be, prescriptive in nature (Borman et al. 2003).

Expanding on the assertion that central authorities are better suited to implement policy development and implementation, further research has discussed the inclusion of other educational tiers that need to be involved in the change process. Honig's (2009) analysis of research into educational policy implementation identifies a growing consensus regarding an implementation model that allows the initial intention defined by a policy to assimilate to the state – district – school level, and thus respond to the specific and contextual variations more effectively. Such a model recognises that the transition of policy from design to practice is a process chain that inevitably involves a transformation of purpose between the original intention and eventual implementation (Adams Jr, 1994; Hamann & Lane, 2004; Honig, 2009; McLaughlin, 1987). Honig (2009, p.337) refers to articles by Datnow (2006) and Malen (2006) who argue that sustainable implementation depends on complicity of action amongst the various actors along the chain that should, in effect, reduce the degree of variation that takes place.

Working along similar lines, Squire & Reigeluth (2000) explore the concept of systemic change and identify four different connotations of the term related to state or national level, district

level, school level, and the last they refer to as ecological. They argue that each of these is dependent on the context of application for its meaning. Work by Smith & O'Day (1990) describes the first of these four contexts and argues that reform policy at the state level implicitly drives change at the local level (Squire & Reigeluth, 2000, p.143). *Teddlie & Stringfield (2007, p.135)* have stated that *"Although policies may set direction and provide a framework for change, they cannot determine outcomes. Implementation tends to predict gains in student achievement.".* Consequently, it has been argued that, although centrally-driven development of reform policy improves the chances of implementation success, *"co-construction"* (Datnow, 2006) of policy and establishing implementation support structures throughout the educational hierarchy, strengthen the sustainability of the reform (Honig, 2009; Smith & O'Day, 1990). It is further posited that this is an effect of limiting prospective cross-tier variations, which in turn should lead to more effective outcomes that are better aligned to the original intentions of the reform policy.

Within the context of this research, the focus remains on systemic change at the state/national level but does not investigate the connections to the school-level tiers directly, choosing to determine if national-level policy change did permeate the system by looking for an overall effect on student outcomes. It should also be noted that arguments about trans-tier coordination made above were recognised in the policy documents with statements being made by officials at the Ministry of Education during the period of policy introduction. These statements reflected a clear understanding on the part of the authorities that policy alone was not enough (Galea, 1999, 2004; Mizzi, 1999). Moreover, these decision makers had recognised that implementation procedures required essential tie-in from across all stakeholder levels through effective change management considerations as reflected in the messages by the director-general for education in each of the policy documents.

> *"The contribution of such a wide spectrum of stakeholders on its formulation was invaluable. Even more crucial will be the commitment and consensus of the same participants during the stage of implementation." (Mizzi, 1999)*

> *"The direction of a strong central authority to monitor development plans and to audit progress cannot however be underestimated. Networking, on the other hand, whilst allowing each school to hold to its identity, mission and ethos, gives strength to initiatives." (Borg, 2004)*

Central authorities, therefore, hold a pivotal role in structuring large-scale reform and systemic change across an educational landscape through policy development and implementation. In leveraging policy to bring about change to education systems, the probability of positively impacting student performance tends to be greater when the change initiatives are introduced

and supported by centralised authorities as opposed to those introduced as more localised initiatives. Cooperation of all actors and stakeholders across all levels of the educational network, together with implementation support structures that are properly established, reinforces the sustainability of the change process. Conversely, a lack of such cooperation or integrated support systems will reduce the sustainability and effectiveness of any change initiative.

2.3 Measuring improvement to monitor reform - A role for Educational Effectiveness Research (EER)

Having established that reform policy needs to be underpinned by reliable evidence is linked to the development phase of such initiatives. This section argues, however, that the sustainability of such reform efforts can only be achieved if integrated with oversight and monitoring mechanisms linked to effective feedback and adjustment systems. Such systems are crucial during the implementation phase to ascertain that introduced policies are working towards achieving their goals and improving the quality of education.

2.3.1 Sustaining effective reform

Without effective implementation, a reform initiative would have little to no impact on school systems or student learning. Bezzina (2003) and Darling-Hammond (2010) posit that educational reform is normally established on the premise of improvement and sustainable positive development of an educational system. Similar arguments are put forward by Fullan (1993), Ginsburg, Cooper, Raghu, & Zegarra (1990), Sahlberg (2010) and MEYE (1999, pp. 2–3, 2004, pp. xi–xii) with the general argument being that such reforms tend to be linked to cultural, societal, political or economic change within a state or region. Although this intended "change for the better" may not be perceived by all stakeholders as "constructive change" (Ginsburg et al. 1990, p.476), the premise of positive development is what usually underpins efforts of measured systemic change in functions within an educational system. These efforts may take place at a micro, meso or macro level[1], and will vary in scope and scale accordingly.

Additionally, the initial stages of the reform planning require an evidence-based understanding of any shortcomings within the system. Understanding systems is key to understanding systemic change and any associated processes (Duffy et al. 2006; Hopkins et al. 2014; Joseph & Reigeluth, 2010; Watson et al. 2008). The challenge is therefore to design reform that will

[1] In the context of this study Stevenson's (2003) definitions are used: Macro-policy refers to policy development at the level of the nation state … Micro-policy is the term applied to policy development at the level of individual institutions, with some commentators identifying an intermediate level, meso-policy at the level of local or regional government (2003, p. 10).

assess and respond to specific needs informed both by an understanding of school effectiveness and school improvement.

Sammons (2009) points out that in response to the emergence of international comparative assessment programs, policy makers and educators around the globe are taking a more robust interest in reform that promotes *"school improvement and improving educational quality and raising educational standards"* (2009, p. 123). Each of these actions lends itself to the "change for the better" premise associated with educational reform and has led to a growing demand for approaches to school improvement that are grounded in effectiveness research (Borman et al. 2003; Creemers & Kyriakides, 2007). The implication here is that designing and implementing educational reform efforts, on any level, requires that the effort be underpinned by an understanding of educational effectiveness research related to both school effectiveness and school improvement, what Louis (2010) refers to as *"a link between the research and the practice"*.

Similarly, Datnow & Park (2010) posit that modern educational policy processes have moved toward systems that create a systemic infrastructure that supports broad-scale change across multiple schools simultaneously. These infrastructures establish a framework within which to deliver reform even though they might vary according to purpose, scope, and context (Lewis & Hogan, 2019, p. 4). Individual studies and papers considering policy-driven reform situations and associated implementation strategies, like those briefly outlined by Duffy et al. (2006); Frick, Thompson, & Koh (2015); Levin (2010); Reigeluth (2006b) (2006a), have some common trends reflected in their considered reform processes and relatively common or similar phases associated with sustaining the reform process. In general, once the reform design is structured, rolled out and implemented — in other words, once the template of school effectiveness is prepared and laid across the school environ with the purpose of adjusting that domain — there is then a switch in focus towards the intended outcomes of the initiative, i.e., guiding the reform towards school improvement. This focused drive is sustained and complemented through established targets, associated monitoring and feedback systems, and integrated support systems that reinforce effective implementation.

2.3.2 Monitoring reform effectiveness

A review of the literature about educational reform processes suggests that effective and sustainable reform needs to be linked to data-driven monitoring and adjustment during the implementation phase. However, such monitoring, although understood to be essential, is not always comprehensive enough to determine effectiveness or impact.

The review by Hopkins et al. (2014), discussed earlier, maintained that education improvement initiatives have shifted from the school level to national/regional level, however, they also propose that an investigation of such systems is what will allow a better understanding of any systemic change. They argue that although educational environments vary as a result of different contexts, an understanding of systemic reform and the characteristics that are common to *"high performing national and regional educational systems"* (2014, p. 271) will not only enable the customisation of the reform effort but also improve the process outcomes.

Datnow & Park (2010); Duffy et al. (2006); Joseph & Reigeluth (2010); Watson et al. (2008) also make the case that emphasises the importance of understanding systemic reform that is driven through a centrally controlled or mandated structure of developmental policies. This position is further reflected in other school reform and improvement literature and is considered to be multidimensional and context-dependent in the general manner in which it is implemented (Hopkins et al. 2014; Townsend, 2007). There are some common threads of what constitutes effective reform that are reflected in the discourse, and these outline the need for: stakeholder involvement in the reform process; strong leadership; highly effective professional development; clarity of objectives and directions for teachers; and the importance of monitoring the effectiveness through data-driven information systems.

Kreber & Brook (2010, p. 96) argue that this last point tends to be lacking in most reform situations so that data-informed oversight and decision making does not play as effective a role as it should in the implementation process. They also point out that although valued, systematic evaluations of reform impact generally tend to focus on more qualitative work analysing participation and satisfaction rates while more quantifying measures of impact are much less prominent.

As discussed in the previous section, the concept of educational reform is itself purposed to improve systems — directly or indirectly — in a manner that delivers positive effects and a better quality of education for all learners. In considering school improvement systems, Bezzina (2003) has argued that *"The key goal of all documents is to improve the quality of education for all students"* (2003, p. 3). Similarly, Supovitz & Taylor (2005) have stated that the purpose of systemic change in education is to ultimately improve student learning. The extent of any change — positive or negative — resulting from the introduction of new policies into an educational landscape tends to be influenced by, and reflected in, a variety of factors. Areas specifically targeted by the policy discourse, but also those not explicitly defined, are all impacted to varying degrees. Creemers & Kyriakides (2007) and Feldhoff & Radisch (2021, p.

9) highlight the complex interplay between different affecting factors and consider "*effectiveness factors as multidimensional constructs*" (Creemers & Kyriakides, 2007, p. 351) requiring dynamic analytical frameworks to understand and render a true picture of what is happening.

Although the systemic reform process broadly affects a complex multitude of different domains and elements within an educational system, the usual goals tend to aspire to an improved and better system. Depending on the context and specific characteristics of the introduced reform, measuring these projected improvements will vary in scope and methodology. Ideally, however, such monitoring is introduced and managed alongside the policy roll-outs as data collection and analysis, and recognised as a key factor to sustain the effectiveness of the reform (Bogotch et al. 2007; Healey & DeStefano, 1997; Hopkins et al. 2014; Sebba, 2003).

2.3.3 Implementation support structures

Successful implementation of large-scale policy reform therefore goes hand in glove with comprehensive monitoring systems that measure variations in a variety of factors and gauge effectiveness and impact. This section reviews the nature of those systems and suggests that the ideal mechanism for monitoring effectiveness and measuring impact is through longitudinal methods that analyse medium to long-term data.

Bogotch et al. (2007) and Sebba (2003) both argue that in order to function effectively and sustain policy-driven reform efforts, the associated support structures need to integrate appropriate monitoring systems that use medium to long-term data collection and analysis to inform on progress and implementation. This essential empirical link to educational effectiveness research is reflected in large national and regional studies conducted by MacBeath (2007); Sackney (2007); Valenzuela, Bellei, & Allende (2016) and stipulated as a reflection on *"…academic performance of an educational system…"* (Viennet & Pont, 2017). However, according to Borman et al. (2003), there tends to be a time lag between the implementation of reform efforts and the subsequent research that explores the impact and outcomes of the reform.

Healey & DeStefano (1997) and Hopkins et al. (2014) have discussed the importance of having timely data to underpin informed decision making during the implementation phase of a reform process and argued that it is a characteristic common to highly effective educational systems. Healey & DeStefano (1997, p. 9) also state that such implementation and support structures, in particular, require the establishment of standards and associated metrics as well

as an accountability environment in order to function effectively. Similar arguments presented by Viennet & Pont (2017) agree with Healey & DeStefano (1997), and consider an integrated and effective data management and analysis system as an essential part of a reform framework. They further argue (also along similar lines as Healey & DeStefano (1997)) that such systems lend themselves to monitoring and evaluation mechanisms and allow a system to reorient itself and make adjustments during the implementation process.

The nature of the data required to support monitoring and evaluation mechanisms can vary depending on the implementation tier at which the policy is functioning (Viennet & Pont, 2017, p. 39), however, when looking at the impact on student learning, the ideal data sets tend to be longitudinal achievement data. Creemers & Kyriakides (2007, p.5) argue that educational effectiveness research requires, necessarily, the collection of longitudinal data and analysis of multilevel organisational structure and Gray, Goldstein, & Thomas (2003) postulate that *"More long-term data would enable us … to look at the effects of changes in schools' policies as they unfold …"* (2003, p.5). However, in discussing long-term systemic change, Valenzuela et al. (2016) recognise the challenges presented to educators and policy makers alike when it comes to triggering and sustaining the educational reform process with appropriate data sets. They identify methodological issues and difficulties with obtaining such data sets to be significant hurdles in the educational effectiveness and improvement research. Bogotch et al. (2007) point to similar *"methodological limitations whether in terms of sampling, designs or statistical analysis"* (2007, p.95). Likewise, Gray et al. (2003) note that longitudinal methods requiring certain techniques for evaluating education effectiveness, such as time series analysis, are dependent on the availability of long series of data covering a span of years. Conversely, if the data series was available and accessible, then the evaluation of education effectiveness using longitudinal analysis would be an appropriate method to apply in determining the effect of a reform initiative.

Popham (1999) recognises that standardised achievement test scores are what authorities and stakeholders use as a measure of school effectiveness, even though, he argues, they are not a valid measure of educational quality. Similarly, Veas et al. (2017, p. 534) recognise that "*Standardised achievement tests are used to provide objective, reliable, and valid measures, with greater use in the field of educational evaluation on a large scale".* Furthermore, Popham (1999) maintains that although standardised tests cannot test all the content within a subject domain, they can be structured to render a *"valid norm-referenced interpretations of a student's status regarding a substantial chunk of content"* (1999, p. 2). This suggests that

achievement test scores can give a rough estimate with respect to the content domain covered by the test construct.

The use of results data drawn from a longitudinal series of standardised tests can therefore be employed as an implementation support structure and interpreted against predefined performance indicators to inform an analysis of progress against the desired or projected outcomes. On their own, these interpretations are not enough to deliver a complete and comprehensive picture of any resulting impact of the entire reform initiative. However, they can be leveraged to determine if and to what extent the reform actions have influenced student learning as determined through standardised testing as along as the domain and test constructs remain the same.

2.3.4 Systemic reform process in Malta

The introduction of the two major educational policies across the Maltese islands in 2000 and 2005 established a policy-driven development context that was intended, in part, to have a measurable effect on the quality of learning and consequently on learning outcomes. The policies in question — NMC and FACTS (MEYE, 1999, 2004a) — were established to give structure to the reform and change processes in Maltese schools over the established periods (MEYE, 1999, 2004a). As such, the policies also acted to guide the reform process that was intended to lead to, among other things, improved quality of education in all its aspects (Galea, 1999, 2004; Mizzi, 1999, 2004). This was therefore a government-led reform process that worked to create a systemic infrastructure to support the intended reform and deliver a greater quality of education for all.

Each of these policy documents established a framework of action to support systemic change in the Maltese educational landscape, a framework that would underpin the change initiative and lead to the desired improved outcomes for learners. In doing so, the documents recognised the importance of educational effectiveness research and monitoring of the reform process *"…so that the true impact of the proposed measures on the educational landscape is objectively gauged." (Mizzi, 1999, p. 5).*

The subsequent reform policy document, FACTS (MEYE, 2004a) similarly established a centralised monitoring and guiding infrastructure to monitor and support the reform initiative.

> *One aspect concerns policy development and co-ordination, standard setting, monitoring and quality audit of the experience and performance of students in all State, Church and private schools. (MEYE, 2004a, p. 29)*

In defining the framework for implementing the intended educational reforms, the FACTS policy established (in law) the Malta Education Directorate (MED) with the responsibility "*to be a quality and standard setter which ensures quality education for all and sponsors good practices…"* (MEYE, 2004a, p. 30). This established the systemic infrastructure required to support the broad-scale changes across multiple schools simultaneously as discussed by Datnow & Park (2010). However, what is less clear is the monitoring systems that were put in place to determine the impact and effects of any reform efforts. Despite stated intention in the policies themselves, there has been little in terms of quantitative analysis of the effect of either policy on student learning or outcomes. This study looks to address the paucity of research in this area by applying a framework for evaluation established around structures designed to universally promote quality of education.

2.4 Framework for structured evaluation

Educational reform, therefore, remains an effort to improve educational function in such a manner as to improve the quality of educational processes and outputs. Such efforts are applicable to any or all tiers of an educational system. The arguments presented in this section suggest that such concepts associated with quality education can provide a foundation on which to construct an analytical framework for understanding change by linking variations across different domains. The framework itself is structured around *Input – Process – Output – Context* domains described by Scheerens et al. (2011a) and to a similar extent Astin & Antonio (2012).

2.4.1 International research efforts

Large-scale national and international research efforts are established in part by monitoring developmental variations in education by gauging changes in quality standards. The Organisation for Economic Co-operation and Development (OECD) and the United Nations Educational, Scientific and Cultural Organization (UNESCO) are two major international players influencing the direction of educational research and development across the globe. The OECD has had an immense impact on governmental policy and strategic planning, in part through their international assessment frameworks (e.g. PISA, TIMMS) (Breakspear, 2012; Sahlberg, 2017), while UNESCO has funded international research in their quest to alleviate poverty through their efforts to improve education (UNESCO, 1990, 2000; Benavot & UNESCO, 2015). Whilst the former entity has funded research into comparing educational provision, mainly across its developed member countries, the focus of UNESCO has been on establishing systems to monitor and improve the quality of education in developing nations.

Both institutions support and conduct major research into educational quality and effectiveness according to their own exigencies and agendas, and both institutions are concerned with researching and establishing systems of education that work on national levels. Although each organisation caters to a different "client base" there are a few common goals towards which each of them works, most prominent of which — as it relates to this study — is a drive to support and improve the quality of education in their member states (OECD, 2013, p. 13, 2018, 2012; UNESCO, 2000, 2005, 2015).

Similar large-scale international research initiatives have worked towards understanding the role and influence of quality education across different nation-states and setting developmental drives with targeted outcomes and observable indicators that would allow a tangible change in the quality of education (American Institutes for Research, 1999; World Bank, n.d.). In most cases, the purpose of establishing the research and development initiative is based on an association between quality education and improved social and economic standards in those countries (Auld et al. 2019; Hanushek & Wößmann, 2007).

> *"…education quality is defined by its contribution to the development of cognitive skills and behavioral traits … that are judged necessary for good citizenship and effective life in the community...the quality of education, has a statistically significant and important positive economic ...(World Bank, n.d.)*

UNESCO's work has been more detailed and considerably more focused on raising educational quality in developing nations with the scope (or hope) of driving development and alleviating poverty (UNESCO, 2015). Their understanding that quality education delivered to as broad a swathe of society as possible (especially to younger members of that society), affects positive pressure on societal and economic factors, that resonates with the World Bank and OECD understanding.

2.4.2 Educational quality frameworks

The use of the term quality has become synonymous with modern-day educational research, reform and delivery efforts associated with improving educational outcomes in almost any contextual setting: *"to deliver quality education", "to improve the quality of education", "for quality education", "quality education for all"* (Benavot & UNESCO, 2015; Borg & Giordmaina, 2012; Hargreaves et al. 2014; OECD, 2013; UNESCO, 2002, 2005). The two main policies that are the subject of this research (MEYE, 1999, 2004a) and the subsequent policy (MEDE, 2012) are replete with references to quality education underpinning the policy initiatives and impacting outcomes for learners.

Quality in education is, however, considered to be a complex concept, contextually dependent, and having meanings contingent on its purpose of use (Harvey & Green, 1993; Iwu et al. 2018; Sahney et al. 2008; Scheerens et al. 2011a). According to Scheerens et al. (2011a), however, it does require more specific definitions when applied to empirical or analytical contexts. Even the hierarchical levels of educational institutions at which the term quality is being employed influence the interpretation of the term though the entire institutional cohort may be working towards a common set of outcomes (Scheerens et al. 2011b, p. 4; Stephens, 2003). This suggests that governance-level or management-level or teacher-level personnel would apply the term to their own contexts in a way that is relevant to their responsibilities.

More relevant to the context of this study, however, is the work done by the larger institutional entities like the OECD, UNESCO and the American Institutes for Research in developing quality frameworks to monitor and/or measure the quality of education by establishing comparative indicators for inputs, process and outcomes (Astin & Antonio, 2012; Scheerens, 2004; Scheerens et al. 2011a). These frameworks are usually associated with, or interpreted against, the relevant contextual factors. This structure is a common framework of reference for comparing, interpreting, measuring, or discussing educational quality.

2.4.3 UNESCO documents for quality frameworks

The concept for applying a quality framework to underpin this study was influenced by the UNESCO frameworks and reports that have applied such systems across different international contexts to monitor impact of developmental programmes. A key characteristic is the ability to link input, process, output, and context variables.

UNESCO's goal to achieve Education for All (EFA) has led them to an understanding that the quality of the education being delivered, in all the aspects and contexts where quality is relevant, is a necessary criterion that must inform any analysis of inputs, process and outcomes.

Education for All: The Dakar Framework for Action (Fiske & UNESCO, 2000) (EFA) was the successor to the 1990 Jomtien World Declaration on Education for All (UNESCO, 1990) that established a global commitment for the global societies and their institutions to work towards ensuring basic education for all and to meet basic learning needs within their communities. King (2007, p. 379) summarised the vision at Jomtien as covering *"…early childhood education, primary schooling, adult literacy, essential skills for youth and adults, and access to knowledge and skills via the mass media.",* a vision that was reduced in the subsequent Dakar initiative.

The initial intent established at Jomtien had not included explicit terms establishing parameters for acceptable quality and standards of the education that was supposed to be delivered. This clarification came with the Dakar framework (UNESCO, 2000), established a decade later after reviews and realisations indicated that having educative structures alone did not necessarily imply proper education was taking place,

> *"Evidence over the past decade has shown that efforts to expand enrolment must be accompanied by attempts to enhance educational quality if children are to be attracted to school, stay there and achieve meaningful learning outcomes" (UNESCO, 2000, p.17).*

The EFA acknowledged, in no uncertain terms, the fundamental link between quality and education, *"The quality of learning is and must be at the heart of EFA"* (UNESCO, 2000, p.21). What the Dakar framework also did was enshrine this intent to work towards quality education by embedding the linkage as one of the six principal goals for all global communities to work towards. Furthermore, Strategy 11 (2000, p.21) of the Dakar Framework for Action established a need for systematic processes to monitor the quality and progress of the different dimensions that exist in the educative world: teaching, learning, resources, environment, community.

Strategy 11 of the EFA concerned itself entirely with the need to monitor quality and progress and stipulated that *"Robust and reliable education statistics, disaggregated and based on accurate census data, are essential if progress is to be properly measured…"* (UNESCO, 2000, p.21) and goes on to encourage governments to continue to develop such capacity

> *"…to produce accurate and timely data, qualitative and quantitative, for analysis and feed-back to policy-makers and practitioners... to identify areas of greatest inequity and to provide data for local-level planning, management and evaluation…" (2000, p.21).*

Subsequent EFA reports continued to improve on the work from Jomtien and Dakar with the 2002 report *"Is The World On Track?"* (UNESCO, 2002) proposing a structured Input-Process-Output framework for defining quality at different levels of an educational hierarchy (2002, p. 80). The 2005 report *"The Quality Imperative"* (UNESCO, 2005) applied the same framework for *"understanding, monitoring and improving education quality"* (2005, p. 35). EFA was succeeded in 2015 by UNESCO's "Education 2030", which reflected a deeper understanding of the implication of the term "quality education" and identified minimum standards of education quality (Auld et al. 2019). It also retained the quality frameworks established in

previous forums and reports with a couple of additional dimensions that would consider the quality of educational design and content (UNESCO, 2015).

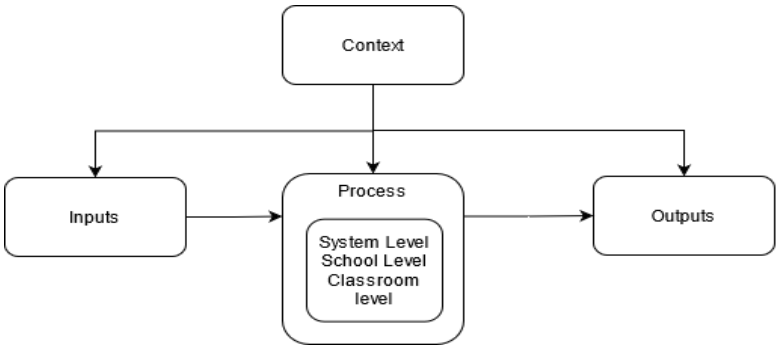
2.4.4 Establishing frameworks for a structured evaluation

UNESCO's "Education for All" and "Education 2030" forums, and their discussion on measuring and monitoring quality education, have therefore been drawn on to inform a key element for organising this study. In particular, the 2002 and 2005 UNESCO "Education for All" forums define a framework by which to determine quality education, namely the *input* – *process* – *output* – *context* framework (i-p-o-c) (UNESCO, 2002) & (UNESCO, 2005). This framework was kept in the subsequent "Education 2030" with explicit recognition of the multidimensional nature of educational quality and the subsequent addition of two other dimensions related to the quality of design and the quality of content.

> *"Monitoring quality in education requires a multidimensional approach covering system design, inputs, content, processes and outcomes."(UNESCO, 2015)*

In reference to the earlier framework design, Scheerens, Luyten, & van Ravens (2011, p.1) have pointed out that *"this framework can be used to clarify a broad range of quality interpretations: productivity, effectiveness, efficiency, responsiveness, equity…".* Scheerens et al. (2011) also emphasise that although this framework has been criticised as being narrow, linear, and authoritarian in nature, it still has broad applicability when it comes to measuring quality education. In this regard, the framework allows flexibility to distinguish and define specific educational functions being investigated and the level/domain (i.e., input – process – output – context) with context underpinning the other three levels. They go on to state that, once established, this framework will facilitate an understanding of any links between the effectiveness at the output domain and variations administered at the other domains.

Figure 2-1: Scheerens et al. (2011a, p. 36)



A similar framework is outlined by Astin & Antonio (2012, p. 19) and uses an *input-environment-output* framework to assess impact. Earlier versions of this model have been

applied as an educational impact assessment framework: *"Astin's model articulates the 3 key elements (I-E-O) of educational impact as previously described and conveys a strategy for analyzing these elements."* (Bird, Anderson, Anaya, & Moore, 2005, p.366).

There is other work by Creemers & Kyriakides (2006) that supports the effective relationship between the different domains. Creemers & Kyriakides (2006) dynamic model of EER presents a complex, multidimensional, multilevel model to describe the interplay of affective educational factors in education. This seems to imply that due to the interrelationship between these factors, making changes to one or several of them results in an overall effect on outcomes. They further posit that this "…*model ultimately explains why educational systems perform differently"* (Creemers & Kyriakides, 2006, p. 352).

In the context of the quality framework and in recognising the multidimensional nature associated with understanding quality in education it is possible to establish a variety of possible systems to monitor, evaluate and compare different levels of quality for different educational domains across different contexts. Creemers & Kyriakides (2006) have also argued that "*It becomes evident from these studies that it is possible to measure a broad range of outcomes in a valid and reliable way using traditional methods of assessment."* (2006)

In order to sustain policy-driven reform and monitor effectiveness, such a framework can therefore be structured to establish correlational, if not causal, links between what is being implemented through the policy and the effects (outcome) reflected in student achievement in the short to medium term, and attainment over longer periods of time. In the case of a broad-reaching national reform policy that targets a broad swathe of educational establishment domains, the overall influence is exerted on the whole of the educational landscape, a "tidal force" across all aspects of the system. The magnitude of such a reform may vary, but the pressure to impart positive influence remains and due to the interconnectedness of different educational domains (Creemers & Kyriakides, 2006), should lead to measurable changes in different aspects of that educational environment. Some parts may react positively and others negatively or remain unchanged, but purposeful actions should result in measurable reactions, changes to inputs or processes or contexts should affect outcomes.

2.4.5 Assessing achievement

The analysis being proposed for this study will draw on a prolonged sequence of standardised assessment results and use them to investigate longitudinal and cross-sectional variations in achievement, and subsequently use these as a measure of outputs. It is therefore appropriate

to briefly review aspects and characteristics of such assessments that would offer objectivity and reliability to the data for such a purpose.

Gerberich's (1963) historical review of educational testing recognises that the need to measure certain human traits or characteristics has always been fundamentally tied to concepts of testing and assessment. Comparing acquired behaviour, knowledge and skill has, inevitably, led to the establishment of assessment tools as a means of measuring or valuing proficiency. His appraisal of testing processes links the measurement of acquired behaviour to achievement test: "*historical antecedents of modern performance tests, oral tests, and written examinations were all designed to measure acquired behaviour, and therefore may be classified as early achievement tests." (Gerberich, 1963, p.185).* As Gerberich goes on to trace the progression of assessment and testing through the early centuries and the later decades, he articulates an evolution of their function that was underpinned by an understanding that objectivity needed to be maintained if their results and associated measurements of achievement were to be meaningful.

Gerberich (1963, p.188) outlines how the measurement of learning outcomes using assessment based on the intended instructional outcomes began around the 1930s which led to the structuring of a system that uses assessment tools to measure variations in student proficiency due to instruction.

Bowers (1991) looks at the factors that influenced the evolution of assessment and testing in the United States of America, with much of the development taking place through the nineteen sixties. Bowers (1991) cites three main factors that had an impact on educational testing, amongst which was a move towards accountability (Bowers, 1991, p.52).

Towards the end of the 1940s, established a framework that used assessment linked to "*vetted goals and objectives*" (Baker, 2013, p.85) to measure achievement. This framework was "*key in an iterative educational process*" (2013, p.85) allowing the system to monitor sequenced learning. This system also introduced more objectivity to the process and enabled the application of assessment systems to a wider cohort through improved standardisation of intended outcomes and associated assessment tools. In this manner, the framework facilitated meaningful cross-sectional and longitudinal comparatives offering an analytical function that would impact our understanding of quality and effectiveness. Baker (2013) also notes that the system has changed very little as a framework that is structured to support educational improvement and defines current usage as a means for determining achievement in either absolute or relative terms (Baker, 2013, p.84).

The arguments made by Baker (2013, p.84), and relevant to the context of this study, state that assessment outcomes are mostly about measuring achievement and are administered usually as different forms of *"products, performances, or processes"* at the school or state level. In administering state-level assessments, there is an effort to maintain assessment standards over short, medium, and long terms for progression and comparative purposes. However, according to Newton (1997) and Patrick (1996), this is tentative over the short term and the validity of such a comparison rapidly drops to irrelevant over longer time spans. Newton (1997) and Patrick's (1996) arguments posit that even in the case of parallel test forms administered over several years, there will be external factors influencing the environment around the different student cohorts that subsequently effect their perceptions and interactions with the test forms. The exact scale or nature of such external influences would, however, be difficult to quantify accurately.

Conversely, however, if the changes in external factors can be shown to be limited, then the different sets of achievement results should offer a better degree of comparability. The general review of the literature indicates that achievement measurements through national standardised tests, over the medium to long term, can therefore be applied to determine if the performance of subsequent student cohorts varies over time. The literature further stipulates that this can be done as long as the test constructs remain strongly comparable and external influences show only slight variation. It remains important though, that consideration be given to those external factors when drawing interpretations, and a more detailed discussion of those elements will be presented in the next chapter.

Ongoing large-scale education reforms are normally structured to change the input, process, and context domains. These changes are designed to have an impact on quality of education and subsequently will affect different areas of the output domain, in particular, achievement results from standardised tests. This study is structured to investigate longitudinal variations in those achievement results to reflect on any impact brought about by the broad-scale policy introductions in 2000 and 2005.

2.5 Chapter summary — Educational effectiveness and quality

The premise that improving the quality of education informs the purpose of most reform actions underpins the entire scope of this study. There is compelling evidence that implementation of such reform processes has a higher probability of success when developed and administered through a centralised authority. There are, however, factors associated with more defined requirements for "effectiveness, equity and quality in education" (OECD, 2013,

p.13) that require establishing evidence-based processes to monitor and measure progress. Defining sets of indicators based on desired goals or projected outcomes strengthens the central authority's capacity to sustain effective, systemic reform and direct evidence-informed decisions accordingly. To this end, governments and education decision makers are making more regular use of both assessments and evaluations to monitor any impacts on student performance and the quality of education.

This chapter has presented arguments that the application of a quality framework will help frame reform actions and support the analysis of any associated impact. The framework outlines four key domains (input-process-output-context) on which such actions are expected to act. More importantly, however, this framework represents, in its simplicity, the complex, multidimensional field of education and intrinsic connectivity between those same domains. Any reform process, intended and planned, will inevitably have an influence on any or all of the domains. Small-scale changes can affect one or more of the domains to varying degrees, but larger-scale systemic reform efforts will, more often than not, be designed to impact the entire educational landscape. Due to the scale of such efforts, any impetus imparted on the system is likely to have an impact on outcomes irrespective of which domains are affected or to what degree.

The NMC and FACTS policies (MEYE, 1999, 2004) introduced different reform drives in their own respect and were designed to improve the quality of education across the Maltese islands. Each was purposed to change a multitude of different domains within the education system as a whole, including an intended purpose to improve student achievement and attainment. This was not the primary purpose of the reforms, rather the main goals were on a broader cultural and social scale for the NMC and a move to decentralisation with FACTS. However, within each of these policies were statements that looked at achievement outcomes associated with improving academic performance and associated achievement and attainment results.

> *Another important issue refers to ascertaining objectively standards achieved by students in the course of their educational experience. (MEYE, 2004a, p. 43)*

The review of associated literature that analysed the impact of these reforms on student outcomes has shown that there is a further need for investigating the effect of the introduced policies. Particularly, there is a need for further analysis of the impact and a clearer measure of the effect that the reform actions may have had on student performance as it pertains to

the cognitive domains of education and more specifically to achievement and attainment on standardised tests.

The focus of the research remains on finding out if the changes posited by large-scale education policies in Malta have had a measurable effect on learning outcomes as immediate and tangible student achievement. Consequently, has each major reform process had an impact on education and learning reflected in student achievement and attainment and subsequently on the overall quality of education across the islands?

3 Constructs, forms, and difficulty levels.

3.1 Chapter overview

The analysis for this study is expected to work primarily with achievement scores and results of students undertaking a common yearly examination. The second chapter of this review considers, therefore, literature associated with test standards, constructs and forms that link the examinations over the stipulated period to factors that could affect outcomes through intended or inadvertent modifications to successive test forms.

The first two sections of this chapter consider how examination standards tend to vary over time and the literature associated with making valid longitudinal comparisons that can support the process of long-term analysis of the examinations in question and their associated student outcomes.

The third section gives an overview of test constructs and associated forms with particular consideration of construct validity and continuity, parallel test forms, and factors that affect test form complexity. Test form complexity will play a part in analysing longitudinal variation in test forms to determine what changes may have taken place before determining how those changes may have affected difficulty levels.

The fourth section looks at cognitive load theory and considers how the mental load experienced by students is related to test constructs and forms. The review of these topics will better inform an analysis of variations in the test forms over the years and help considerations of whether any changes to student achievement resulted from shifts in learning standards, or modifications to paper setting, or both.

In considering variations in difficulty levels of test forms, the final section reviews a statistical mechanism in the form of item analysis that will be used in this research to render a

determination of any tangible change in the quality of the examinations. This particular analytical pathway is established on a well-structured collection of records associated with the Junior Lyceum examinations from 1998 to 2010 (Curriculum Department & Educational Assessment Unit, 1998 - 2010) that offered a consistent annual set of item analysis data.

3.2 Examination standards over time

In the context of analysing student examination outcomes over prolonged periods of time, the concepts and discourse associated with the manner in which examination standards change is both relevant and necessary to compare longitudinal data for the same test constructs. Such discourse and associated research and literature will inform interpretations of the outcomes and underpin inferences drawn from the analysis. The principal issue here is that simply looking at pass-fail rates over time, detached and devoid of context, is by no means an absolute indicator or holistic measure of improvement in teaching and learning standards — nor is there any definitive way in which to measure such standards absolutely (Crisp & Novaković, 2009, p.4; Fitz-Gibbon & Vincent, 1997; Jones & Ratcliffe, 1996; Newton, 2005).

Patrick (1996) and Newton (1997) have argued that long term longitudinal comparisons of educational standards lose meaning as the comparative time span increases due to changing contexts associated with changing culture and technology (Coe et al. 2008). As mentioned in the previous chapter, Newton (1997) and Patrick (1996) have stipulated that a sequence of parallel test forms based on the same, or very similar test constructs, would not necessarily represent an unchanging continuity in standards and difficulty levels for consecutive student cohorts.

On the other hand, there is a recognition that although comparability loses significance with time, shorter periods would offer more meaningful comparisons. Crisp & Novaković (2009) have stated that there is more significance to an analysis when considered over shorter periods of time, and Newton (1997, p.227) also recognises that *“The notion of applying the same standard' becomes more and more meaningless the further apart the comparison years”.*

A further point by Newton (1997), — linked to this notion of longitudinal comparisons losing significance over longer periods — maintains that the complexity of influences on the effectiveness of education, and the associated stimuli impacting education quality and standards are extensive. Analogous to Heraclitus’ statement *"You cannot step into the same river twice, for other waters are continually flowing on.",* Newton's (1997) arguments suggests that examination contexts tend to be dynamic, and each situation will be unique. Furthermore, the intricacy of any causal analysis and associated interpretation is just as

complex, and fluctuating pass-fail rates are merely a reflection of some change in one or several associated factors (Crisp & Novaković, 2009; Newton, 1997).

However, an argument made by Kane (2013) noted that interpretation of test scores can be used for "*policy analysis, program evaluation, research, and educational accountability*" as long as the construct validity is maintained. Coe (2010), Goldstein & Cresswell (1996) and Kolen & Brennan (2014) have all argued that as long as there exists a common construct linking examinations together, and appropriate consideration is given to all the relevant factors, then comparisons can be legitimately made. If a time series analysis of the achievement results were to be conducted over a span of several years, then any fluctuations in standards, cognitive loads, and difficulty level characteristics of the test forms need to be considered and, where possible, factored into the analysis.

The general understanding seems to hold, therefore, that reliability of meaningful conclusions drawn from a comparative analysis of examination outcomes — for examinations established on the same construct — has an inverse relationship to time. As such, if this argument is taken to underpin the research, then comparing exams with similar constructs over shorter periods of a few years should render an acceptably reliable comparative analysis over the short term from which to draw more meaningful interpretations.

3.3 What to compare? What to measure?

This research is concerned with identifying and, if possible, measuring changes in education quality and school improvement systems resulting from national policy changes and reflected in changing student assessment outcomes over time. The relationships between policy change, learning standards and student outcomes are by no means straightforward, however, as argued earlier, if there is general progress in teaching and learning systems as a result of policy change, then such an improvement should be reflected somewhat in improved outcomes and display some form of evidence of impact.

As discussed in the previous chapter, there is a multidimensional reality to the effectiveness of policy implementation and those changes introduced by larger organisations in a top-down manner tend to be more successful, but less accurate in predicting outcomes due to transformation of purpose (Adams Jr, 1994; Hamann & Lane, 2004; Honig, 2009; McLaughlin, 1987; Teddlie & Stringfield, 2007). The multidimensional change factors act to influence outcomes in general and the broad scale of implementation of a national policy therefore acts to create a situation whereby no one factor can be definitively linked to one particular measurement of effect. Ball (1993, p. 15) recognises that for situations where there are

multiple reform initiatives through a policy process, or the introduction of a multitude of grouped or sequenced policies, there is usually a cumulative effect on general educational domains and associated outcomes.

Although broad-scale national policy changes, like those introduced in Malta in 2000 and 2005, work to influence a multitude of different educational aspects, they also tend to create micro-level variations that are dependent on institutional contexts (Lewis & Hogan, 2019). The general outcome results in a collective effect on teaching and learning across the educational landscape. According to Cohen & Hill (2001) such policy introductions tend to have a range of different stimuli depending on implementation strategies in schools and classrooms.

In order to determine the general effect of these large-scale policy introductions, there needs to be an assessment of impact at the macro-level and spanning the whole educational landscape. This would require developing an exploratory and investigative evaluation framework to look at general trends in school improvement indicators associated with the implemented policy changes. This approach is what Cohen & Hill (2001, p.184) describe as a "Black Box" situation looking at the overall effect rather than considering differentiated processes.

Effectively, such an evaluation will take the form of an analysis of variations in student outcomes reflected in changing examination results over time, underpinned by an unmodified set of success criteria. The study will compare one aspect of policy impact on student outcomes through a review of student results in successive sittings of the Junior Lyceum Entrance Examination[2] (JLEE)— more specifically, an analysis of variation in pass-fail rates; aggregated grade averages; and proportional grade distributions. The analysis will be based on five annual exams (which constitute the whole examination set), described by the MEYE (1999, p.59) and MEYE. (2004, p.43) as being established on the same educational construct for each of five subjects.

Although this approach only looks at one educational tier — primary education attainment — it does have broad relevance in understanding the impact on Maltese pre-tertiary education in general due to the interconnectedness of the Maltese educational system. These examinations were recognised as a core national examination benchmark (Curriculum Department & Educational Assessment Unit, 2005, p. v), and were administered to a significant

[2] The Junior Lyceum Examination was a qualifying examination, coordinated by the Education Division, taken at the end of the primary school course. All those who pass qualify for entry into a Junior Lyceum. The others continue their education in Area Secondary Schools. This examination is optional. (MEYE., 2004, p.43)

portion of the Year six population — 72% on average — during the same period that the two major policies under consideration were introduced.

3.4 Test constructs and test forms

In order to sustain comparative relevance, constructs informing subsequent test forms need to be continuous or relatively similar (Coe, 2010; Newton, 2005). Additionally, subsequent test forms based on the same construct need to be consistent in standard of complexity to minimise any influence from varying difficulty levels on overall outcomes (Crisp & Novaković, 2009). Understanding the degree of parallelism along the series of JLEE examinations is the subject of this section together with a review of related literature with the scope of informing analytical structures discussed in the subsequent chapters.

The test forms for the Junior Lyceum Examinations were coordinated and prepared centrally (MEYE, 1999, 2004a) which according to Grima et al. (2008, p.29) were intended to be a qualifying examination at the end of Year 6. As such, they were developed to a common construct — designed to assess the same content and with corresponding statistical specifications (Kolen & Brennan, 2014) — for each of the subjects (Grima et al. 2008). Nevertheless, even for test forms developed according to a common construct one issue that persists from one examination to the next is a variation, however slight, in complexity and associated difficulty levels that tend to be objectively intangible and defy precise control (Alberts, 2001; Coe, 2010; Kolen & Brennan, 2014). In this regard, an analysis of continuity and consistency of the test constructs, and variations in structure, complexity, and difficulty levels of the associated test forms, can be considered to determine if any such variations impacted outcomes. The suggestion being proposed at this point is to apply a longitudinal analysis of these factors allowing the analysis to incorporate their possible effects in drawing conclusions.

The premise considers that determining the degree of association and variation between successive test forms for these two factors will inform an understanding of how parallel the test forms remained over time. Depending on the degree of similarity from one year to the next or, for that matter over longer periods, determinations can be made as to what degree of influence on student outcomes may have been a result of changes in test form items. If the similarity is strong enough to consider test forms parallel, then it becomes possible to associate any change in outcomes with other influencing factors. If not, then the tests will have no basis for comparison within the context described.

3.4.1 Linking constructs and comparing outcomes

According to Chapelle (1998), constructs are taken to be inferences regarding elements of a learner's particular competence, skill or knowledge (Feuer et al. 1998, p. 12), setting a framework used to define what is to be measured by a test form (Irwing & Hughes, 2018). Furthermore, Coe (2010, p.279) and Newton (2005, p.107) argue that test forms can only be effectively compared to each other as long as there are common linking constructs that can be used to establish a comparative context, a view also held by Dorans et al. (2010) and Kolen & Brennan (2014).

Newton (2005, p.109) stipulates that the greater the similarity between the constructs, the better is the association and therefore the comparison. Moreover, Coe (2010) further proposes that this can also be extended and considered for different subjects as long as there are common constructs that can be used to associate the test forms. Reasoning along similar lines, Kolen & Brennan (2014) discuss designs that assess the same content with corresponding statistical specifications, establishing a comparative framework within which to analyse outcomes. Veas et al. (2020) applied similar construct-based comparisons to determine inter-subject correlation for an academic performance construct.

Further arguments by Newton (2005) posit that if a common construct can be established between different test forms, *"designed to assess the same construct in the same way"* (2005, p.107), then it can be used to define a common frame of reference within which those test forms are associated or linked. This frame of reference can be used to support a comparative analysis of student outcomes from one test year to the next. As suggested earlier, such a comparative would work to inform a year-on-year analysis of variations in outcomes and, if extended to more prolonged periods of time, would reflect on possible variations in trends.

Nonetheless, Fitz-Gibbon & Vincent (1997) and Jones & Ratcliffe (1996) have pointed out that although the analysis can be done within a common frame of reference, defined by the linking constructs, there will still be a multitude of other influences that will have a slight impact on the outcomes. This understanding resonates with arguments put forward in the previous section that large-scale reform policies affected a multitude of different domains and sub-domains across the educational landscape to varying degrees. As a consequence, any lines of association determined by the construct can only establish similarity – imitation not duplication.

A case in point would be different comprehension passages selected for successive English language tests can be chosen to be similar but will be drawn from different texts and the effect

of the nuances of each text on outcomes cannot be categorically explained. Much of the literature discussing this issue (Coe, 2010; Crisp & Novaković, 2009; Dorans et al. 2010; Newton, 1997; Patrick, 1996) offer similar arguments to the effect that: the variations introduced when using different question items while trying to retain the same assessment context and construct, cannot realistically be considered in their entirety. Rush et al. (2016) consider the effect of writing flaws and item complexity on difficulty and discrimination indices, arguing that test writing standards need to be maintained to deliver continuity on successive test forms.

However, a detailed impact assessment of each of the slight effects that different questions may have on students, or the slight variations in students' learning experiences from one year to the next that may marginally modify response patterns, is beyond any scope or purpose of this study. The arguments put forward by Feuer et al. (1998) and Cohen & Hill (2001, p.184), argue that the comparative does not need to dive into the micro-level analysis to assess year-on-year variations, rather it would be more effective to consider a more robust macro-level analysis of general influences (planned or unintended) on outcomes within the different test forms themselves.

Relevant to the linking of test constructs considered here, Dorans, Pommerich, & Holland (2007), Feuer et al. (1998) and Newton (1997, 2005) all consider the process of equating as a statistical mechanism designed to bring the scores from different tests forms onto the same scale. Similarly, work by Dorans et al. (2010) and Kolen & Brennan (2014) states that this process eliminates the variations in scores due to inadvertent fluctuations in difficulty levels. Direct use of equating mechanisms will not, however, be used as part of the analysis of data as there are no raw scores available to process. Nevertheless, a brief consideration of the principles of equating is indispensable as they are the same general principles for linking test forms established on the same constructs.

3.4.1.1 Equating

Equating is a process that adjusts for differences in difficulty among forms that are built to be similar in difficulty and content (Dorans et al. 2007, 2010; Kolen & Brennan, 2014). As long as the test forms have been developed to test the same construct — with the same content and statistical specifications — equating can be used to compare the tests and draw inferences (Alberts, 2001; Dorans et al. 2010; Feuer et al. 1998; Newton, 2005).

Horizontal test equating can be used to equate tests administered to the same Year level in consecutive years. Dorans et al. (2010, 2007) and Livingston (2004) elaborate on a process

that requires equating raw scores from one test form to another set of scores established by a "base" or "reference" form. Similar processes for equating have been described by Hanson (1993) and Kolen & Brennan (2014). Furthermore, the principles of test equating assume the results data is continuous rather than discrete. Equating can still, however, be applied to discrete test results using a system of kernelling (Dorans et al. 2007; Hanson, 1993; Kolen & Brennan, 2014).

The process of equating is applied as a transform function to the raw scores on a new test form relative to the base form (Dorans et al. 2010; Goldstein & Cresswell, 1996). Once the transform is applied, the scores can be scaled accordingly and uniform comparisons across the test forms can be made (Livingston, 2004). The literature on the topic of equating does however make an effort to emphasise the limitations of equating and the caution that needs to be taken in drawing absolute conclusions from the results (Dorans et al. 2007; Livingston, 2004; Newton, 1997, 2005).

The scores data from the Junior Lyceum Examinations used for this research are established in discrete scale levels which themselves cannot be equated efficiently as the scales cover five broad score bands from grade level A to grade level E. The associated raw scores are not readily available for analysis. So, although the use of equating test forms across the different years would have rendered a reliable idea of any variations in difficulty reflected in the equating function, the lack of availability of raw scores to undertake this procedure will not allow such a process to be employed effectively.

Linking of test forms will instead need to be established on a comparative array that will look to identify any longitudinal variations in the test construct to determine the level of association and similarity from year to year and, if possible, over a longer period of time. This comparison will need to be supported through a construct validation process.

3.4.2 Construct validity – Content and marking schemes analysis

In the absence of equating mechanisms, it becomes necessary to establish a system that uses construct validity to establish longitudinal links between test forms. Construct validity asserts that the construct underpinning a set of similar test forms establishes the context for comparison and linking of those test forms and validates any interpretations or inferences made thereof (American Educational Research Association et al. 2014; Coe, 2010; Downing, 2003; Feuer et al. 1998; Kane, 2013; Kolen & Brennan, 2014). It has also been argued by Downing (2003) and Kane (2013) that such validity is not a property of the test or test forms, but of the interpretations of the test results.

According to Kolen & Brennan (2014, p. 487) the process of equating is purposed *"to put scores from two or more tests on the same scale—in some sense"*. Feuer et al. (1998, p.15) have a similar understanding on the concept of linking in assessments, and both are referring to establishing a commonality for comparing test forms through an equating process. The underlying principle of linking test forms does however remain so that "like" can be compared to "like", and a comparative array can alternatively inform such an evaluation (Feuer et al. 1998, p.15).

There is therefore, a need for validation of any underlying test parallelism that will underpin comparative studies. This, together with an understanding of the degree to which a series of test forms maintains consistency and alignment with the established construct will go some way towards establishing a validity argument (Chapelle, 1998; Coe, 2010; Downing, 2003; Kane, 2013). Coe (2010) discusses construct validity and interpretation of examination results and advances the notion that if comparing of results from different examinations is shown to be valid then the construct must be common and can be used to interpret "*the results of those examinations in terms of that construct.*" (2010, p. 279). He goes on to say that an instrument of two or more examinations may be used to determine construct validity in interpreting results by understanding the content and determining if it renders an *"internally consistent measure"* (2010, p. 280). Similar work done by Veas et al. (2020) used *"academic performance"* as a common construct for comparability, recognising that inter-subject comparability on this basis alone may be considered tenuous (2020, p. 78), but possible. Coe (2010) and Veas et al.'s (2020) arguments centre on more general educational concepts than those defined by a singular subject domain, and can be considered a *"special case of construct comparability"* (Coe, 2010, p.279) that consider parallel test forms.

So, whether considering inter- or intra- subject construct comparability, validation of the common construct in terms of coherence and continuity is essential to support interpretations and inferences made through the analysis and use of the test scores (Kane, 2013). Downing (2003) looks at construct validity and having an associated and coherent validity argument as the only key feature required to underpin interpretations of test results. Similar arguments are presented by Kane (2013) who references the American Educational Research Association, et al. (1985) as the standard establishing that *"validation is the overall evaluation of score interpretation"* (Kane, 2013, p.7). Both Downing (2003) and Kane (2013) further state that a validity argument needs to be established through multiple sources of evidence to "*support or refute meaningful score interpretation*" (Downing, 2003, p.831).

The issue for this study therefore becomes establishing a relevant construct validity argument. Work by Kolen & Brennan (2014) proposes that this can be done around an analysis of content and associated statistical specifications, each a principal factor defining a testing construct. Downing (2003) presents a table of items that can be used as possible sources of validity evidence categorising them as evidence types associated with *"content, response process, internal structure, relationship to other variables, and consequences"* (2003, p. 832). Three of these categories, 'content', 'response process' and 'consequences', are more relevant in the context of this study and the JLEE examinations and can be informed using the available documentation and data sets. Furthermore, analysis of types of evidence associated with "*content"* and "*response process"* will also be applicable to other areas of the data analysis, in particular, analysis of test form complexity and psychometric characteristics, while "*analysis of consequences"* resonates with the intended policy analysis.

The other categories of validity evidence described by Downing (2003, p.832) are contingent on the validity argument and not crucial at this time due to the parallel nature of the examinations and their purpose — to measure similar cohorts in a similar context through similar sets of examinations. "*Internal structures"* will technically be considered indirectly through an analysis of complexity and difficulty levels. However, "*relationship to other variables"*, which deals with correlations and statistical relationships to other forms of measure (Downing, 2003, p. 835), is not considered essential as there are no other external measures against which to compare.

This section is therefore proposing that longitudinal continuity and consistency for a set of JLEE examinations can be investigated by comparing the constructs over the period in question in order to determine the degree of association through a validity argument. In light of the data available to this study, comparison of content and associated marking schemes can be applied to validate construct continuity (Coe, 2010; Kane, 2013). Such an analysis will look at commonalities and differences across the sets of test forms to understand construct consistency over the years. The next two sub-sections elaborate a little more on these two areas (content analysis and scoring characteristics) to inform the analytical framework being proposed for this study and support associated construct validity arguments.

3.4.2.1 Content Analysis

Content analysis is one type of evidence used to support validity arguments (Chapelle, 1998, p.49; Downing, 2003) and, according to Krippendorff (1989), help inform deeper inferences

that are not immediately obvious in the text. Work by Moeini (2020) and Shaw & Crisp (2012), present similar arguments linking validity and inferences based on test scores.

As stated in the previous section, the validation process will substantiate the degree of consistency and continuity of examination standards across the period in question through an analysis of content items on each of the tests — a determination of the examination blueprint (American Educational Research Association et al. 2014; Downing, 2003). This breakdown of the papers will allow a measure of the degree of consistency and identify any variations in the longitudinal trend. Any such variations could then be considered for their effect on the overall achievement and results of the student cohort or explain possible changes in a time series.

However, further consideration regarding other affecting factors is needed. Chapelle's (1998) discussion regarding content analysis as part of a validity argument sees an understanding of operational constructions and associated differences for test-takers as important factors for consideration. In particular, whether or not there were any variations to the general structures of the exam item formats and sequences that may have influenced contextual factors and, as a result, the subjects' operational constructions (Chapelle, 1998, p.53).

Considering the data available for analysis and in relation to the categories tabulated by Downing (2003, p.832), the associated sets of possible types of evidence that can be investigated and applied to a validation argument would need to centre around an analysis of "content" and "response process". This can be applied longitudinally to the different test forms to determine trend variations and the degree of parallelism over the years.

3.4.2.2 Scoring characteristics

Scoring characteristics can be considered a subcategory of content analysis and test specifications (American Educational Research Association et al. 2014, p.14; Downing, 2003, p.832), however, the consideration here is to assign more particular attention to assigned score weightings of the topical groups for each subject and their items or subtopics. The intention is to determine if there was any reweighting or redistribution of scores over time.

A comparative analysis would therefore need to determine to what extent the scoring, associated marks distribution and associated marking schemes remained similar over the years. Parallel test forms would be expected to have parallel (or quasi-parallel) scoring schedules and marking schemes. Feuer et al. (1998) argue that as part of the linking construct, a scoring schedule can be given particular attention as it is straightforward to compare marks distribution between test forms especially if those forms are very similar. This can primarily be

used to support any validity argument and conclusions about the degree of similarity between test forms and the degree of continuity of the construct over the years.

3.4.3 Test form complexity – Affecting student outcomes

As stated earlier, although there can be no guarantees of exactness in calculating the difference in relative difficulty, it is nevertheless possible to determine and consider fluctuations in difficulty levels in different ways to get a better understanding of any resulting effects on student outcomes.

Feuer et al. (1998, p.15) maintain that even when there are no deliberate changes to a test construct from one year to the next, there will still be slight differences that impact difficulty levels in subtle ways. This assertion is backed by other works by Crisp & Novaković (2009) and Dorans et al. (2010) who also argue that there will always be small differences that impact difficulty levels in small ways despite best efforts to maintain uniformity from one test form to the next. These differences exist as minor variations in such things as cohort characteristics, demographics, teaching and learning practices, curricular emphasis and slight discrepancies in the way tests are written, scored and marked (Coe, 2010; Kolen & Brennan, 2014; Newton, 1997; Patrick, 1996). All these factors, together with other small changes, play a part in slightly affecting the characteristics of both the student cohort and the test forms consequently having an overall effect on the relative difficulty levels for that sitting.

Consider, for instance, slight changes that teachers make to their teaching from year to year — sometimes in reaction to a previous examination question on a particular topic, or as part of their own professional development — these changes take place as teachers mature personally and professionally and impact student learning to some extent. The same holds true for student cohorts that are influenced by their changing environments which may be minor on a year-to-year basis but are there nonetheless. Similarly, curricular emphasis may shift slightly from one year to the next, transmitted by such things as formal curricular notices or examination reports being used as feedback for schools to act. These influencing factors tend to be localised to class groups, year groups or schools and not a uniformly shared stimulus equally distributed to the entire student cohort. Furthermore, the earlier discussion regarding large-scale policy implementation implied that different schools or regions implement policy changes inconsistently and centralised support works to manage the convergence towards common goals (Healey & DeStefano, 1997; Hopkins et al. 2014; Nunnery, 1998). As such, any resulting impact on relative difficulty level would not necessarily be uniform across the entire student cohort.

On the other hand, there are other subtle changes to the test form itself that can have a more uniform, direct influence on all students being examined. The commonality of the examination to all those sitting the examination makes variations to the test form itself, a uniform change common to all. These are seen as small variations to the way the test form was written, richer or additional graphics to support text, variations in text readability, changes to the way problems are presented on the paper (Hewitt & Homan, 2003; Lee & Heyworth, 2000; Luo & Skehan, 2007). Crisp & Novaković (2009, p.4) argue that these factors have an effect on the overall difficulty level of a test form, albeit to varying degrees, and consequently impact pass-fail rates.

Although all these influencing factors act together as a whole to modify the relative difficulty level of a test form creating a unique educative context for each subsequent cohort, the latter set of factors, those directly embedded in the examination papers, can prove to be more tangible and observable as factors equally affecting an entire student cohort. These small changes in subsequent test forms can consequently be considered and compared. In more precise terms, a comparative analysis can be made for the more observable, common changes that are likely to have an impact on outcomes to better understand any associated effects (Jones & Ratcliffe, 1996).

Considering that difficulty levels are taken into consideration as test form developers try to maintain uniform levels of difficulty from one test form to the next, there is no way of guaranteeing absolute homogeneity, as the concept of difficulty is a relative construct (Coe, 2010; Crisp & Novaković, 2009; Fitz-Gibbon & Vincent, 1997; Kolen & Brennan, 2014; Newton, 1997). Newton (2005) maintains that although examination boards have attempted to maintain consistency in linking standards for sequential test forms, the reality is that there are limitations to the mechanisms for linking and on how true these links are.

Notable and relevant in all of this is that year on year differences, whether intended or not, can result in sizeable variations to these examination standards over prolonged periods. Moreover, these variations inadvertently do have an impact on outcomes of test forms designed to the same construct specifications and applied to similar cohorts over a relatively short time span. So, while the relative difficulty levels of similar test forms can be approximated using similar content and statistical specifications, there can be no guarantees of exactness and the consequences will be reflected in the variance between pass-fail rates from one year to the next.

This has particular relevance in the context of this research that considers a set of five exit exams, repeated year on year for similar Year level cohorts (Year six students), and intentionally designed to the same test constructs. A comparison of these successive test forms will need to inform the investigation into policy impact on student outcomes by considering their alignment to their common construct and determining the degree to which successive test forms have remained parallel.

3.4.4 Cognitive and statistical analysis of test form complexity

Investigating the relative variations in complexity and difficulty levels reflected in the test forms' structures and text will shed light on any changes in the mental load exerted by the exam papers over the years and add another dimension to interpreting policy impact. Recognising the link between test complexity and difficulty levels, this section proposes the use of both cognitive and psychometric measurements as effective mechanisms to determine variations in test form difficulty.

Although test forms may be parallel and devised according to the same construct, this does not mean that the individual items of assessment reflect the same complexity, textual levels of facility, or ease of comprehension across the different forms. Beckmann et al. (2017) and Pelánek et al. (2022) describe the complexity of a task as an intrinsic property defined internal structures, while difficulty level is more subjective and relates to the interaction of the student with the task itself. Furthermore, Beckmann et al. (2017, p. 2) argue for a distinction between the two concepts, describing task complexity as a *"cognitive concept"* and item difficulty as a *"psychometric concept"*.

However, according to Sax, Eilenberg, & Klockars (1972) there is some level of correlation between item complexity and item difficulty (as determined by difficulty ratios). Similarly, Lee & Heyworth's (2000) discussion of problem complexity and difficulty recognises that more complex items are considered to be more difficult. Brindley (1987), Candlin (1993) and Nunan & Keobke (1995) make similar links between task complexity and associated difficulty levels. It is therefore being posited here that for different test forms prepared for the same test construct, the exact correlational relationship between task complexity and difficulty levels will not need to be considered in any particular detail. These can be considered to be proportionally related such that increased task complexity implies an increased difficulty level, allowing the study to determine variations in each separately with the overall analysis of both informing an understanding of any variation in difficulty level of the exams.

It should be noted that according to Coe et al. (2008), Dorans et al. (2010) and Fitz-Gibbon & Vincent (1997), getting an absolute and objective measure of the true difficulty level of a test form and any calculated differences between subsequent forms, although important, is not conclusively possible in a practical manner. Following from the literature discussed in the preceding sections, even if an accurate formulation were applied that associated all items across two similar test forms — allowing an exact measure of difficulty level and thus an exact calculation of variance — once the test forms are applied to different cohorts, other factors are introduced making such alignments tentative at best. Influencing factors like these would be relative variables associated with cohort characteristics that are not easy to define or quantify in their totality and with accuracy (Paas et al. 2003; Paas & Van Merriënboer, 1994).

Nonetheless, in order to compare student performance outcomes longitudinally, a general understanding of change in the complexity and relative difficulty of linked test forms is required to interpret the results accordingly — any large variations in either would be expected to impact student performance outcomes. However, due to the complexity of defining an exact evaluation of difficulty levels it should, on the other hand, be possible to establish a comparative appraisal of the forms and render a directional estimate – more difficult, less difficult, same difficulty — associated with an approximation of the magnitude of difference. Such a vector would, as Newton (1997) and Patrick (1996) pointed out, be relevant to subsequent test forms in the short term, which in this case may be able to shed light on any general trends over the longer term.

One possibility to consider here is a comparison of test forms over the years with particular attention to an analysis of psychometric and cognitive variation in difficulty levels as discussed by Newman et al. (1988). A statistical analysis of facility and discrimination indices calculated for test items on subsequent tests can be drawn from the same data recorded at the time for each test form for each sitting from 1999 to 2010 (Curriculum Department & Educational Assessment Unit, 1999 - 2010). The cognitive analysis, on the other hand, would need to consider variations in one or more of the following item characteristics: cognitive demand (Downing, 2003; Hancock, 1994; Jones et al. 2009); readability (Chapelle, 1998; Gillmor et al. 2015; Hewitt & Homan, 2003); and extraneous load factors (Clark et al. 2011; Gillmor et al. 2015).

In the retrospective context of this study, both statistical and cognitive methodologies can be drawn on to inform the understanding of variations in difficulty of the test forms over time, with the latter method being associated with the field of Cognitive Load Theory (Bannert,

2002; DeLeeuw & Mayer, 2008; Paas & Van Merriënboer, 1994; Paas et al. 2003; Sweller, 1988, 1994) and the former being a statistical item analysis based on outcomes (Karelia et al. 2013; Matlock-Hetzel, 1997; Sim & Rasiah, 2006). The possibilities offered by using both these methodologies lies not in establishing a precise measure of changing complexity or difficulty level, but in determining overall longitudinal trends, if any, in test form complexity and the impact on student outcomes.

3.5 Cognitive Load Theory

Cognitive Load Theory (CLT) (Sweller, 1988) offers the possibility to objectively interpret test forms and their associated task items in terms of complexity and difficulty levels and thus offers the opportunity to determine comparative degrees of variance between test forms. There has however been a scarcity of research that considers linking CLT to assessment (Gillmor et al. 2015; Kettler et al. 2009) and much less, if any, that attempt to use CLT to interpret the complexity of a test form.

Nonetheless, the prospects offered by CLT can support an interpretation of changes in the complexity of test forms over time and consequently associated fluctuations in difficulty levels as discussed in the previous section. This can be achieved by comparing the intrinsic properties of a test construct over time to determine any degree of change, and also through a comparative analysis of the extraneous factors on the associated test forms (such as readability, number of process steps in a mathematical problem, cognitive level of action verbs used, paper settings, number of visual aids, etc. (Gillmor et al. 2015)). The hypothesis being proposed here is that such an avenue of investigation will lend support to the study by offering another layer of analysis that would enhance understanding of variations in complexity. That will further inform interpretations of any fluctuations in student outcomes over the same period associated with a variance in cognitive load (CL) due to variations in the test construct (intrinsic CL factors), or test forms (extraneous CL factors), or both.

Krell (2017) and Paas & Van Merriënboer (1994) recognise CL to be a multidimensional construct characterising the effort required by a learner to answer or respond to a particular task, with more challenging tasks considered to have a higher CL than less challenging tasks. The multidimensional nature of CLT is reflected in the categorisation of factors that influence the CL experienced by a test subject: intrinsic cognitive load, extraneous cognitive load and germane cognitive load (DeLeeuw & Mayer, 2008; Gillmor et al. 2015; Paas et al. 2003; Paas & Van Merriënboer, 1994; Sweller, 1988, 1994; Sweller et al. 1998).

Intrinsic cognitive load depends on the complexity of the constructs being worked on (DeLeeuw & Mayer, 2008) and is a function of the number of elements being processed and their level of interactivity. "*An element is anything that has been or needs to be learned, most frequently a schema*." (Sweller et al. 1998, p.259). According to Kettler et al. (2009) it is therefore an intrinsic property of the constructs themselves. Although Sweller et al. (2019, p. 264) also associate the learner's "expertise" as an affecting factor.

Extraneous cognitive load in CLT is considered to be an effect of the way the material is structured and presented to the learner, it is associated with the format of the material (Gillmor et al. 2015; Sweller, 1988). Changing the instructional design for a task, therefore, has a direct effect on the extraneous load exerted by that task (Sweller et al. 2019).

Germane cognitive load is a learner-centred characteristic and depends on such things as previous knowledge, motivation, learning preferences etc. (DeLeeuw & Mayer, 2008; Sweller, 1994; Sweller et al. 1998). Germane CL is subjective and depends on an individual (DeLeeuw & Mayer, 2008; Sweller et al. 1998). Reconsideration by Sweller et al. (2019) has redefined the concept to assume 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)

3.5.1 Linking CLT to assessment

A brief review of CLT literature associated with assessment practices is in order at this stage as it will underpin the analysis of intrinsic and extraneous CLs exerted by the exam papers and, subsequently, a longitudinal comparative of these characteristics.

The concepts and perspectives discussed throughout CLT literature are associated mainly with instructional design and integrate subjective and objective dimensions in considering the context of that design. Sweller et al. (1998, p.263) note, however, that CLs exerted by problem-solving situations are more relevant to testing and assessment contexts rather than learning contexts. Both Gillmor et al. (2015) and Kettler et al. (2009) postulate that due to the strong link between instructional tasks and assessment tasks on test forms, CLT can be similarly applied to both. Beddow (2018) and Kettler et al. (2018) consider test items to have adjustable qualities using principles founded in CLT and describe test accessibility as *"the degree to which a test and its constituent item set permit the test-taker to demonstrate his or her knowledge of the target construct of the test."* (Beddow, 2018, p. 199). This implies that, similar to instructional tasks, the CLs of assessment items have a variable quality to them

which can be modified and controlled for. This implication in turn suggests that such a quality can underpin comparative criteria between test items to objectively determine a relative CL.

From a more investigational point of view, Gillmor et al. (2015, p.4) and Beddow et al. (2008) point out that some recent studies (Gillmor et al. 2015; Kettler et al. 2009; Miller, 2011) have worked to modify test items to vary the CL of those items on the test subjects without modifying the intended construct. These studies worked to understand the effect on a subject's performance as a consequence of redesigning test items to assess the same test construct with modified extraneous load reflected in modified task characteristics (better layout, simplified wording, etc.).

However, in respect to the design for which they were carried out, all three studies — (Gillmor et al. 2015; Kettler et al. 2009; Miller, 2011) — differ from the main thesis being proposed here in their purpose of application. The requirements of this research are associated with using CLT to inform a longitudinal comparison of trends in the complexity of test constructs first, and subsequently of associated test form characteristics. Furthermore, it should be noted that in the context of this research, comparative analysis of these extraneous factors allows broader latitudes of interpretation as it looks for relative variations in these factors rather than attempting to interpret an exact magnitude of difficulty.

3.5.2 Mental load and complexity of a test form

> *"Mental load refers to the load that is imposed by task (environmental) demands. These demands may pertain to task-intrinsic aspects, such as element interactivity, … and to task-extraneous aspects associated with instructional design."(Sweller et al. 1998, p.266)*

In developing curricular materials, Sweller et al. (1998) and Paas & Van Merriënboer (1994) define mental load as the combination of intrinsic and extraneous CLs collectively acting on the test subject. Sweller et al. (1998) postulate that these two factors have an additive effect on the subject. The new conceptualisation of these factors by Sweller et al. (2019) merely reinforces this principle, redefining the concept of germane load as a go-between that redistributes working memory according to intrinsic load (2019, p. 264). Furthermore, Paas & Van Merriënboer (1994, pp. 354–355) establish a link between the mental load of a task and the complexity of the task. The complexity or difficulty of a set task therefore becomes a function of the task characteristics and is dependent on the number of elements or schema used simultaneously within the task and modified by extraneous factors (Beddow, 2018; DeLeeuw & Mayer, 2008; Paas & Van Merriënboer, 1994; Sweller et al. 1998).

It can therefore be postulated that in developing a test construct and associated test forms, the manner in which the test items are structured — i.e., the number of elements that need to be handled by the test subjects; the applied formats; the jargon or vocabulary used; and the level of thinking skills expected — all work to generate an overall CL for that test form. Krell (2017) maintains that this CL interacts with each test subject creating a mental load depending on the more subjective characteristics of both test subject and test form. Complexity can therefore be considered a function of this interaction (Sweller et al. 2019, p. 264) affecting mental load (Krell, 2017).

However, the cognitive loading of the test form is a task-centred characteristic, independent of other more subjective characteristics associated with extraneous and germane loads (DeLeeuw & Mayer, 2008; Paas et al. 2003, p. 65). It therefore offers a potentially objective insight into task complexity and subsequently any associated difficulty. Although this complexity may objectively remain a function of the task characteristics, such a function is not, however, an absolute determination of difficulty levels for projecting outcomes. The influence of subjectivity and germane functions in test-taking still have a major role to play in the reality of those outcomes.

Haladyna & Rodriguez (2013) have argued that *"No item has a natural cognitive demand"* (2013, p. 33), and similar to the discussions of Paas et al. (2003) and Paas & Van Merriënboer (1994), imply that difficulty level can never be determined as a definitive measure of any one test item or test as a whole. Nonetheless, such measures can be expressed as an approximate degree of complexity and associated mental load determined from the various more objective affecting factors that can be identified more readily and quantified accordingly.

3.5.3 Affecting factors

In working to determine trends in test form complexity and consequently trends in difficulty levels, it becomes necessary to establish a comparative framework of affecting mental load factors to be compared. In the context of this study the fact that the data is collected from past reports makes an analysis of germane factors an insurmountable challenge, however, intrinsic, and extraneous factors can be given appropriate consideration due to the availability for analysis of the examination reports and the examination papers.

In considering the intrinsic and extraneous factors for subsequent test forms, this study will want to identify those factors that lend themselves to comparative analysis. Comparing test constructs will reflect on any changes in the intrinsic load factors over time while comparing test form variation will shed light on extraneous load factors.

3.5.3.1 Relevant intrinsic factors

Kettler et al. (2009, pp.539–540) argue that "*the items on a test should demand only those cognitive resources intrinsic to the target constructs they are intended to measure*". They go on to say that affecting extraneous and germane factors should be eliminated in designing test forms as they hamper the subject's *"capacity to demonstrate performance on the target construct"* (2009, p. 539). Similar definitions are put forward by American Educational Research Association et al. (2014) and reasserted by Beddow (2018).

The brunt of the mental load of an assessment construct should therefore be carried by the target constructs alone, and these should establish the defining foundation on which to structure the overall assessment construct while minimising the impact of extraneous and germane loads (American Educational Research Association et al. 2014; Beddow, 2018; Kettler et al. 2009). These arguments, together with the considerations discussed by Coe (2010) Kolen & Brennan (2014) and Newton (2005) regarding the conditions for linking constructs for comparative outcomes, suggest that any comparative validity of test outcomes from successive tests would require longitudinal continuity. The emphasis on the intrinsic affecting factors thus furthers the earlier arguments regarding the necessary continuity of an assessment construct to support the comparative analysis underpinning this research.

In reviewing the intrinsic factors, emphasis should therefore be put on analysing and comparing any changes in content and statistical specifications set within the test construct and ascertaining the level of continuity or variation accordingly.

3.5.3.2 Relevant extraneous factors

The discussion presented in these last sections suggests that there are several relativistic influences affecting the complexity of a test paper, and any variation to extraneous factors would affect the interaction of the test subjects with the exam (Embretson & Wetzel, 1987; Krell, 2017; Paas & Van Merriënboer, 1994; Sweller et al. 1998, 2019). This section briefly considers the debate around extraneous CL in order to support an analytical framework (based on accessibility) to be used to investigate any associated variations across parallel test forms.

The general understanding in the literature is that the formats and structures of a test form play an important role in establishing the extraneous CL. Work by Hancock (1994) and Melovitz Vasan et al. (2018) has recognised that different assessment formats could measure similar constructs albeit at different cognitive levels — Multiple Choice Question (MCQ) formats, for example, had an impact on reducing the cognitive demand. Similarly Martinez (1999) stipulates that different format test papers exert different cognitive demands on test

subjects, affecting outcomes. Simkin & Kuechler (2005), and later Rudolph et al. (2019), discuss a more granular approach in two different educational domains, investigating test item formats to analyse item-level CLs. Both Rudolph et al. (2019) and Simkin & Kuechler (2005) used Blooms taxonomy and establish a *"knowledge level analysis"* (Simkin & Kuechler, 2005, p. 79).

There is other work by Kettler et al. (2009) and Krell (2017) that links extraneous CL to outcomes that is relevant to this study. Both have argued that in comparing year on year outcomes for a common construct, consideration of change in extraneous factors becomes important in determining a truer picture of any variations in those outcomes.

In order to facilitate an analysis of extraneous CL factors and determine impact on outcomes, this study is therefore proposing the use of an analytical framework that uses concepts of accessibility (Beddow et al. 2008; Kettler et al. 2009, 2018) to investigate test form variations. The underlying principles associated with retaining accessibility (Beddow et al. 2008) by controlling extraneous factors for test takers, is established on an understanding of variation in extraneous CL. Beddow et al. (2008, p. 3) look to reduce excessive CL by moderating for extraneous material, reading load, and the visual impact of the items. Similar work by Kettler et al. (2009) examined extraneous factors that affected students with learning disabilities for whom any variation in difficulty could be amplified by their disability.

A more general approach by Kettler et al. (2009) postulated that in determining the level of accessibility of a test item, the main extraneous factors impacting the mental load of the test taker were affected to some degree by the general format, complexity (reflected in the cognitive demands of the test items) and readability of the items (2009, p. 532). Similarly, Lee & Heyworth (2000) consider *"cognitive variables"*(2000, p. 87) of the applied syntax used in the questioning text, number of steps to arrive at a solution and familiarity of story context as factors affecting difficulty.

In comparing variations of extraneous load factors on a test form, this research therefore requires comparability frameworks that can be utilised effectively by allowing the test forms themselves to be processed and analysed. Kettler et al.'s (2009) determination of accessibility offers a tentative analytical framework based on a content analysis of the readability levels of the presented text passages, the cognitive item demands made by the questions and the general format and structures of the items.

The next three sub-sections discuss these three areas being proposed to structure the analytical framework to support the longitudinal investigating of variations in extraneous CL of the JLEE examinations.

i. Readability

Krippendorff (1989) and Chapelle (1998) consider readability as an important methodology supporting content analysis in education and determining the CL or difficulty level associated with any particular test (Chapelle, 1998, p. 53). Gillmor et al. (2015) worked to directly modify the complexity of test items by changing how they were structured thus allowing the extraneous CL to be varied for the same construct. More specifically, Gillmor et al. (2015) looked at factors impacting readability, format and numeric complexity and made variations accordingly. Meanwhile, Mifsud (2019) has applied readability measures to compare difficulty levels between Maltese and English texts on international examinations delivered to the same cohorts to show that the difficulty levels were not the same. Hewitt & Homan (2003, p. 13) have presented strong correlational data between reading ability and item difficulty which they argue reinforces *"…the importance of item readability as a factor of item difficulty."*

In a paper reviewing what characteristics affect text difficulty for a reader, Anderson & Davison (1986) have argued that although readability formulas can account for some variance in text difficulty level (1986, p. 9), they cannot give a definitive measure of complexity or difficulty level of a text based on statistical properties of the words and sentences alone. Similarly, Reck & Reck (2007) have also recognised that readability scores do have limitations and may not necessarily reflect true text complexity when the calculating equations are dependent on word and sentence length. However, Reck & Reck (2007) have also pointed out that a relative comparison of texts using the same readability tool can be useful for describing texts (2007, p. 1) relative to each other. It is in the longitudinal context of the study that comparative readability measurements become relevant, identifying variations in complexity of the texts from one year to the next.

Most readability formulae and indices tend to be language-dependent and although several are accessible for English, none have been found that were specifically adapted to Maltese. One readability formula that has been tried and tested on Maltese texts by (G. Mifsud, 2019, p. 65) was the Lasbarhetsindex (LIX) formula. There are however two other formulas that may be considered in the Maltese context that are similarly not language-dependent but rely on text statistics. These are the Automated Readability Index (ARI) and the Coleman-Liau Index (CLI). This section is proposing that that all three formulas remain relevant as comparative

mechanisms for Maltese examination texts, while a broader set of readability mechanisms can be used for the English examinations.

ii. Cognitive item demands

Jones et al. (2009) working in the context of higher education sought *to establish the degree of* variation in cognitive levels expected by the intended learning outcomes and associated assessment questions. Similarly, Downing (2003, p. 833) argued that there needs to be a demonstratable link between the curricular cognitive expectations and the cognitive demands integrated on the associated assessment forms. In working to establish the cognitive characteristic of test forms, Brucia (2020, p. 24) and Hancock (1994), have argued that there are a variety of methods for classifying test items by cognitive level. Furthermore, both authors have recognised that Bloom's taxonomy seems to be more prevalent and pragmatic for such purposes.

Jones et al. (2009) leveraged Bloom's taxonomy to determine the *"…difficulty level of each question in the examination paper … from the criteria of keyword/s found in the question."*(2009). Their framework offered a concise mapping mechanism for determining cognitive demands through an analysis of *"examination question verbs"*(2009, p. 3). Their reasoning behind choosing Bloom's Taxonomy, as opposed to other frameworks, was due to this particular taxonomy being i) recognisable and familiar in academic circles; ii) broadly applicable across different subject matter; iii) fairly straightforward to apply due to its simple structure (Jones et al. 2009, p. 1). Similarly, Newman et al. (1988), leveraged Bloom's Taxonomy to categorise item cognitive levels in their research while Chang & Chung (2009) and Dueñas et al. (2015) applied Bloom's taxonomy as part of their algorithmic item analysis to determine cognitive levels of question items automatically.

In the context of this research, this taxonomical framework will support an investigation into trend variations in mental loads exerted by subsequent test forms by underpinning analysis of variations in cognitive demands presented in each of the question items. Objectively determining the exact level of item difficulty is neither plausible nor necessary, however, if the application of Jones et al. (2009) methodologies of cognitive demand classification — higher, intermediate, and lower cognitive levels — is applied, then general variations in the question characteristics can be traced and compared.

iii. General format and structures of the items

Hancock (1994) and Martinez (1999) had recognised that the different assessment formats could measure similar constructs albeit at different cognitive levels, implying that question

format had an impact on the cognitive demand being made. This in turn indicates that as long as the format of the question types remained the same for the same construct, then the associated cognitive demands remained unaffected.

Efforts to change the extraneous load exerted by the various test items will also have had an impact on changing the mental load on test-takers. Clark et al. (2011), have argued that reducing CL on learning materials can be achieved through improved use of visual aids to represent spatial information, appropriate signalling designed to focus attention and minimisation of extraneous visual and textual factors.

Arguing along similar lines, Gillmor et al. (2015) have stated that extraneous CL is also influenced by the general structure and format of assessment items that impact working memory. Their study sought to control for CL by modifying test items to reduce extraneous factors that *"…may contribute to construct-irrelevant variance in order to more accurately measure the intended construct."* (2015, p. 1). Furthermore, they conclude that assessment items are less complex for students when they signal important information, are organisationally easy to follow, and have all extraneous information removed so that they only measure the intended construct.

Work by Miller (2011) looks at aesthetics on the design of e-assessments and the link between cognitive-easing and aesthetical design. However, this main association between Miller's work and the intended analytical framework herein is in line with the ideas posited by Clark et al. (2011) and Gillmor et al. (2015) in looking at the general organisation of the test papers and items. The premise being posited here is that a better organised paper with clear signalling and supporting visuals would facilitate working memory by reducing mental load.

3.5.3.3 Variability of cognitive difficulty level

Readability, cognitive demands exerted by the action verbs, and the general layout and format of test items are three avenues of comparison that need to be reviewed and compared longitudinally to show complexity trends associated with a common test construct over time. These areas will help determine any changes that may have been implemented by the test writers to affect extraneous loads exerted on test takers. Similar to the work by Gillmor et al. (2015) the intention herein is to identify any changes in the CL of test items that may have impacted working memory allowing students to work more or less efficiently. This analytical process will work to understand the cognitive level of variation represented in the subsequent test forms and will need to be processed and presented alongside the statistical analysis of difficulty levels determined from the outcomes.

3.6 Item analysis: Facility and discrimination indices

The analysis of variation of cognitive levels and mental load exerted by the examination papers on the different student cohorts is considered part of the contextual analysis being used for this study. As discussed earlier, cross-referencing this information with psychometric measurements from each examination should work to better inform a longitudinal trend analysis of test complexity. The continuous computation of Facility and Discrimination indices by the EAU between 1999 and 2010 (Curriculum Department & Educational Assessment Unit, 1999 - 2010) presented this study with a concise set of data for reviewing these psychometric characteristics. It also offered a means of rendering a longitudinal picture against which to compare the analysis of aspects of the different demands of the examinations.

It needs to be noted that, although considered applicable and useful psychometric indicators of tests and test items, both the facility and discrimination indices are not definitive in their function. Lee & Heyworth (2000) recognised that although the facility index offers an acceptable measure of the difficulty level of items, it is *"not generally agreed that it adequately represents the degree of cognitive challenge an item is to students"* (2000, p. 85). Similarly, Pyrczak (1973) and Joshi et al. (2020) argued for similar caution when considering the associated discrimination indices and their implications. However, review of the literature has associated the discrimination index with the quality of tests and test items suggesting that greater discrimination is one characteristic reflecting better quality items. Chiavaroli & Familari (2011), Doneva et al. (2018) and Pyrczak (1973) applied it to MCQ questions while Azzopardi & Azzopardi (2020), Joshi et al. (2020) and Khoshaim & Rashid (2016) drew similar inferences for non-MCQ questions. Accordingly, it remains important that this type of analysis be applied in conjunction with those discussed in previous sections to establish a more substantiated trend analysis of variations in test form complexity over time.

A review of the relevant literature dealing with facility and discrimination indices of test items will be discussed to elaborate on the definitions and equations relevant to this context and support explanations of how these measures were applied by the EAU and subsequently by this study. Likewise, discussion of the relationship between Discrimination and Facility is then presented as an initial explanation of analytical tools developed in this study to better understand the longitudinal changes in the quality of parallel test forms and associated examination standards.

3.6.1 Defining the facility and discrimination indices

The facility index (F) of a test item is described as the proportion of test-takers who answer the test item correctly (Ebel & Frisbie, 1991; Matlock-Hetzel, 1997; McCowan & McCowan, 1999). This is the generally accepted understanding for determining a quantitative measure of the difficulty level of a test item with Karelia, Pillai, & Vegada (2013) and Sim & Rasiah (2006), describing the difficulty index as the ratio between the total number of correct responses and the total number of responses. In a situation where each question of the test form carries a score of 1 as may be the case with MCQs, then these two definitions are equivalent and easier test items will have a higher value than more difficult test items.

The discrimination index (D_i), or discrimination power (Azzopardi & Azzopardi, 2020; Escudero et al. 2000; Matlock-Hetzel, 1997; Metsämuuronen, 2018), of a test item, which runs on a centre 0 linear scale is a measure of how well the test item differentiates between those students who have the ability to do well in answering the test from those who do not. Metsämuuronen (2018) begins by explaining that the discrimination index is a "loose term" and that a higher D_i score signifies that the test item discriminates more efficiently between students, while lower scores mean that the item is not effective in distinguishing less capable students from their more capable counterparts in the context of that test. Similarly, the Education Assessment Unit has defined the discrimination index as a correlational measure between *"those who score high marks on the item and those who score high marks on the test"* (Curriculum Department & Educational Assessment Unit, 2005, p. 14).

Much of the associated literature review discrimination power as it applies to MCQ tests with test items that tend to be dichotomous in nature and singular in weight (Karelia et al. 2013; Sim & Rasiah, 2006). Work by Jandaghi & Shaterian (2008) has considered the relationships between the variables in terms of weighted scores for each of the test items and established the relationship in such a way as to integrate the proportionality of the respective item results rather than individual singular scores.

3.6.2 Calculating facility and discrimination indices

As part of an item analysis exercise conducted on a test, F and D_i are determined to derive information about how well the test functioned in assessing students by understanding these two psychometric characteristics for each item presented.

In considering F, most item analysis research associated with these statistical measures is based on MCQ tests. However, it remains possible for non-MCQ test items to be processed in a similar manner (Jandaghi & Shaterian, 2008; Khoshaim & Rashid, 2016), albeit with a slight

variation to the calculation. For MCQ test items where each item has a maximum score of 1, F is the ratio of correct responses to the total number of responses.

$$F = \frac{Number\ of\ correct\ responses}{Total\ number\ of\ responses} \quad \ldots\ldots\ldots\ldots\ldots \quad (1)$$

For non-MCQ questions, where the score weight may vary from one item to the next, the ratio becomes that of item average mark to item maximum mark.

$$F = \frac{Average\ score\ on\ item}{Maximum\ score\ of\ item} \quad \ldots\ldots\ldots\ldots\ldots \quad (2)$$

On the other hand, the D_i of a test item is taken as a correlational association between the test item and the test as a whole. According to Matlock-Hetzel (1997) a good test item would reflect a truer probability of a good student succeeding on the test and a lower probability for a weaker student. Conversely, test items with a lower discrimination score have a weaker correlation to the overall test result. In discriminating effectively between those students who can achieve a higher score compared to those who do not, Hotiu (2006) argues that D_i can be taken as a measure of item quality, (2006, p. 24). The association between D_i and the quality of test items is also proposed by Ebel & Frisbie (1991) who define a basic *"…rule of thumb for determining the quality of items with respect to their discrimination index."* (Suruchi & Rana, 2014). DiBattista & Kurzawa (2011) and Musa et al. (2018) have effectively applied analysis of these psychometric measurements to evaluate the quality of MCQ-based medical test items. Azzopardi & Azzopardi (2020) and Jandaghi & Shaterian (2008) have applied discrimination analysis to longer answer type questions that were non-MCQ.

D_i is taken to be the difference between the proportion of correct responses to the item achieved by the upper 27% and the proportion of correct responses for the lower 27% of the cohort.

$$\mathrm{D_i} = \frac{(\mathrm{UG}) - (\mathrm{LG})}{\mathrm{N_{largest\ group}}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (3)$$

D_i = Discrimination index for test item i

UG = Number of correct answers for upper 27%

LG = Number of correct answers for lower 27%

N = Total number of students in whichever group is larger (N_U or N_L)

For non-MCQ items, where item scores can be greater than 1, Jandaghi & Shaterian (2008) proposed a similar calculation to determine the discrimination index of a test item that

considers the proportionality of the test item scores. An adapted version would have the following relationship.

$$D_i = \frac{\Sigma UG_i - \Sigma LG_i}{N_{avg} \times m_i} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (4)$$

D_i = Discrimination index for test item i

ΣUG_i = Sum of scores for upper 27%

ΣLG_i = Sum of scores for lower 27%

N_{avg} = Average number of students in both groups

m_i = Total mark of question i

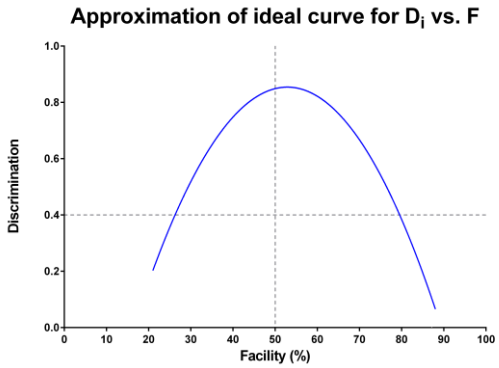
Although there is general agreement on working out D_i , there is however a slight difference of opinion on an ideal cutoff mark that would stipulate a good test item. Karelia et al. (2013) consider a score at or above 0.25 - 0.35, while Hotiu (2006) suggests a score of 0.5, that Sim & Rasiah (2006) associate with moderately difficult test items. On the other hand, the Curriculum Department & Educational Assessment Unit (1999 - 2010) considering an index band around the 0.4 or above (Table 3-1 below). The general consensus is that the requirement will be subjective to the test writers' interpretations; that the D_i should be positive; and D_i should fall in the range between +0.3 and +1.0. It follows that the more test items fall above any such mark the more representative or effective is the test form.

Furthermore, Hotiu (2006) posits that when D_i is approximately 0 then the items are either too difficult or too easy to be discerning enough, while a negative D_i would represent an inverse discrimination and not reflect the true capacity of the test taker for that test also affecting the overall outcomes.

3.6.3 Relationship between facility and discrimination indices

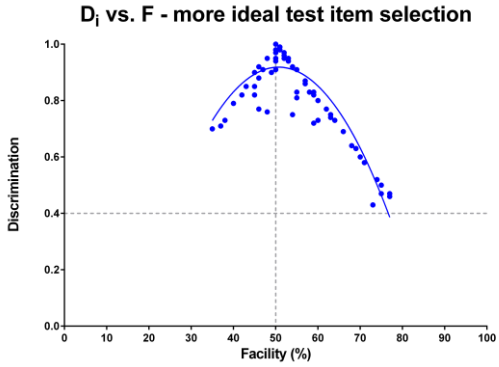
Aiken (1979), Hotiu (2006) and Sim & Rasiah (2006) discussing the relationship between item difficulty and discrimination indices, have argued that such a relationship was ideally non-linear. Sim & Rasiah (2006) further stated that on a plot of Discrimination vs Facility, a "dome-shaped" graph would represent the more difficult and easier test items discriminating less than those question types that have a more moderate difficulty level. Figure 3-1 demonstrates an inverted parabolic shape representing such a distribution and approximates a typical plot for what the literature describes as an "ideal" test with controlled distribution of test items with moderated difficulty levels.

Figure 3-1 Exemplar of an approximation to an ideal curve for a Discrimination vs Facility Plots



Karelia et al. (2013) citing Sim & Rasiah (2006) have argued along similar lines that an ideal test set for assessing effectively would be structured to have a reduced or controlled number of very difficult or easy question types. In other words, referencing the above plot (Figure 3-1), if the item response analysis were to show fewer outlying points and a tighter, higher curve on the same scale, then the balance of facility and discrimination would represent what might be considered a more effective set of test items (Figure 3-2).

Figure 3-2 Plot of discrimination vs Facility for a more ideal set of test items



The exact shape of the curve will, nevertheless, depend on the psychometric characteristics of the test, and those will vary depending on the test design and purpose. Ebel & Frisbie (1991) had stipulated that "How difficult a test should be relates to the purpose for testing and the kind of score interpretation desired.", implying a link between the psychometric attributes and the test construct.

The definition of an ideal test curve at this stage will establish a frame of reference against which to compare and describe. The scatter plots of D_i vs F for large-scale tests can subsequently be used to determine test characteristics at a glance. On a unit scale for each of the axis (D_i range -1 to +1 and F range 0 to 1), the distribution density of the plots can shed light on the facility of the test overall and the practicality of the test in assessing the construct with that cohort.

3.6.4 Junior Lyceum examinations item analysis

The Junior Lyceum examination assessment report prepared by the assessment unit, conducted an item analysis for each exam item on each of the examinations from 1999 to 2010. They calculated the facility and discrimination index for each test item and used this information to report the test characteristics for each year. (Curriculum Department & Educational Assessment Unit, 1999 - 2010).

> *"The facility index of an item is a value that indicates the proportion of students that get the item correct… The facility index for an item is easily calculated by dividing the mean (average) score on the item by the maximum mark that can be scored on the item"*
>
> *"The discrimination index for an item…measures how well the question distinguished between candidates. This is usually found by measuring the correlation (relationship) between the score on the item and the total test score."*
>
> *(Curriculum Department & Educational Assessment Unit, 2005)*

For the facility index, the EAU applied equation (2) above while equation (4) was applied to determine the discrimination index for each test item. The discrimination index was calculated using the following equation:

$$d = p(UG) - p(LG)$$

d = discrimination index

p(UG) = proportion of correct answer for Upper Group (27%)

p(LG) = proportion of correct answer for Lower Group (27%)

Similar equations are proposed by Hotiu, 2006 (p. 24) who considers the difference in terms of proportions.

The Curriculum Department & Educational Assessment Unit (2005)[3] stated that an ideal facility of 0.5 was desirable "*particularly if a question carries a good number of marks (e.g. an essay)"* (2005, p. 15). Although stated, however, no item analysis was ever conducted for essay type questions and this was noted in the reports (2005, p. 16). The EAU further state that a discrimination index above + 0.4 is desired to reflect appropriate test items. For each item response analysis for each exam, the reports presented the following table and analysed the item distribution against it.

[3] The EAU report from 2005 is taken as a reference point, but the text and statements are repeated in meaning or verbatim throughout all the reports spanning 1999 – 2010.

Table 3-1 Levels of Difficulty and Discrimination

Levels of Difficulty and Discrimination
F = 40%-60%; D=0.4 or more (correct levels of difficulty and discrimination)
F= 40%-60%; D = more than 0.3 but less than 0.4 (correct levels of difficulty and discriminated sufficiently)
F = less than 40%; D 0.3 or more (on the difficult side but discriminated sufficiently)
F = more than 60%; D = 0.3 or more (on the easy side but discriminated sufficiently)
F = 40%-60%; D below, 0.3 (F correct but item did not discriminate sufficiently)
F = less than 40%; D = below 0.3 (on the difficult side and did not discriminate sufficiently)
F-more than 60%; D = below 0.3 (on the easy side and did not discriminate sufficiently)

(Curriculum Department & Educational Assessment Unit, 2005, p. 18)

Furthermore, the item analysis was conducted on a sample of 200 test scripts in each of the five subjects. However, as stated earlier, the analysis did not include a few of the more subjective items from the test papers (essays and paragraph responses).

3.7 Chapter summary and implications for the study

The sections in the second chapter are intended to support an understanding of variation in the test forms over the years to consider if there were any changes to student achievement that may have resulted from a variation in examination standards, or differences in difficulty levels, or both.

The first section of chapter two, concerned with understanding how examination standards are affected over time, surmised that the longer the period for analysis between parallel test forms, the less valid are any associative conclusions that can be made between those test forms. Subsequently, the methodology for analysis will need to rely on shorter periods on either side of the policy introduction when interpreting results rather than the longer 14-year period that encompasses the whole of the data set.

The subsequent sections considered linking test constructs and CL theory with the mental loads exerted by test forms. These two sections will help inform how different test forms based on a common test construct will inadvertently have a varying mental load on different cohorts. The literature review on these matters, and in the context of this study, led to a premise that it would be better to inform an analysis of change over longer periods by looking at trend variations in test form difficulty level for successive sittings. These trends could then be used to review any changes in student achievement and results in light of any variations.

The fourth section of this chapter was a review of item analysis literature and set the initial stage for establishing a means of measuring variation in difficulty levels for parallel test forms specific to this study. This last section has led to an understanding that there are three

possibilities that can lead to changes on a graph of discrimination vs facility for the same test construct.

i. The student aptitudes remain the same, but the test complexity changes.
ii. The student aptitude changes (better or worse) but the test complexity remains the same.
iii. Both student aptitude and test complexity change.

In summary, the longitudinal analysis of overall student outcomes from a standardised test set administered over a prolonged period of time needs to be considered in periodical chunks and against an underlying understanding of variations to construct and difficulty levels of those same tests. This literature review establishes a basis to structure a framework for analysis and allows the analytical mechanisms described in the methodological section which follows to be grounded on accepted theories and practices.

Section 2: Research design and methods

Section Overview

The research design and methods used for this study and the underlying rationale are described and discussed over the next three chapters. The methodologies are established to investigate the three key record sets available for analysis: the policy documents (MEYE, 1999, 2004a), the EAU examination reports (Curriculum Department & Educational Assessment Unit, 1997 - 2010), and the record of results (Educational Assessment Unit, 1997–2010).

Chapter 4 gives structure to the research and presents the main aim and associated research questions that will inform the selection of methods that follow. The chapter proceeds to review the different research aspects associated with documentary and data analysis, frames the study's progression and interpretation of the analysis, and discusses the overall methodological framework required.

Chapter 5 develops the methodologies needed to conduct the retroactive longitudinal study and presents critical structures for parallel analysis of year-on-year examination contexts and their respective achievements and outcomes.

The closing chapter of this section acts as a supplementary section to the methodology and is structured to explain the conversion and preparation of data for analysis and give a description of other preliminary processes related to the various analytical methods described in the preceding chapter. To this end, chapter 6 offers a detailed discussion of the digitisation and error-checking procedures needed to convert printed hardcopies to machine-readable softcopies. The data sources available and received as hardcopies included: the record of student achievement results; the item analysis data; and the comprehension texts for Maltese and English.

4 Design

4.1 Chapter overview

The introductory section for this chapter reiterates the context of the research along with the general outline of the organisational framework. The introduction also establishes the research questions (RQ) and presents the rationale behind them before outlining the methodologies needed to address them.

The second section reviews the literature on integrated methodologies and considers the suitability of such methods for this particular study. It establishes the overall analytical design structured on the i-p-o-c quality framework and presents a schematic of the processes applied.

The third part offers a more detailed discussion of the characteristics of the documents and records used in this study and the respective documentary analysis applied. Each set of documentation required a different analytical approach due to their different purpose in the official record.

The last part of this chapter considers the characteristics of the data sets available for analysis as well as the validity and reliability. It also briefly discusses the record of student results as a documented data source (having all student JLEE outcomes from 1997 – 2010) as a prelude to the digitisation process presented in Chapter 6 - Digitisation Process.

4.2 Introduction

Investigating the Junior Lyceum Examination results and associated documentation was intended to facilitate cross-sectional and longitudinal comparisons and determine whether or not the introduction of the NMC and FACTS policies impacted educational quality reflected in student outcomes. In so doing, the study was structured to analyse and understand the

changes in terms of a quality analysis framework and investigated input–process–output—context factors and relationships associated with the large-scale policy introductions. As discussed earlier, such systems can be applied as a framework to underpin structured evaluations (Astin & Antonio, 2012; Scheerens et al. 2011a). The investigation subsequently progressed through three key inquiries into the whole change process — a documentary analysis of the policy changes, a longitudinal analysis of construct validity of the JLEE, and a time series analysis of the associated results data — with a separate investigative method being used for each data source and their interpretations being linked through the quality framework.

In terms of input, the study reviewed the two key policies directly, their discourse and intention, and the manner in which they were structured to affect the process, context, and outputs of the education system in Malta. The analysis of variations in the results was more directly associated with understanding changing outputs in terms of attainment on the sets of Year Six exit examinations. However, as discussed earlier, the comparison of these results over the period during which the policies were introduced would depend in part on the test construct validity over those years (Chapelle, 1998; Coe, 2010; Downing, 2003; Kane, 2013). It therefore became necessary to determine the degree of consistency and continuity of those test constructs and their associated test forms over the fourteen years in question. This latter investigation was an inquiry into the examination as a process and a review of the context reflected through the EAU reports.

Furthermore, longitudinal changes in these aspects of the JLEE would have implications when interpreting variations in the examination outcomes. A fair degree of longitudinal consistency would suggest that variations in student achievement on the JLEE may have been a result of the policies' influence on other domains in the educational landscape. This premise guided the general scope of this research and was the main consideration in formulating the RQs for this study.

4.2.1 Research question

The introduction of two consecutive broad-scale educational policies by the Ministry of Education and Employment in 2000 and 2005 were intended to have a direct effect on learning outcomes and student attainment for the general student cohort (MEYE, 1999, 2004a). The policies introduced paradigm shifts to the educational landscape and were meant, in part, to improve and advance student-centred learning and, through decentralisation of the decision-making processes, empower schools to do the same (MEYE, 1999, 2004a).

However, the extent to which these policies have had an impact on student outcomes has not been explored on a broad scale. This research seeks to inform a general understanding of these policy effects by reviewing and analysing the impact on a fixed set of Year 6 transitional examinations considered to be "end of primary school" benchmarks (Curriculum Department & Educational Assessment Unit, 2005, p. v).

A systematic analysis of the relevant data spanning the introduction of these policies was conducted into associated student achievement and structured to address the following question:

Have the changes introduced by large-scale education policies in Malta had a measurable effect on learning outcomes as immediate and tangible changes to student achievement on the Junior Lyceum Entrance Examinations?

This primary question was subsequently organised into a set of sub-questions that would consider various aspects associated with its primary goal. These were purposed to deliver a structured methodology along with the analytical tools required to inform a considered analysis of the impact on student achievement. Apart from considering the achievement scores, those tools also needed to consider intrinsic and extraneous factors influencing the test constructs and forms and their effect on the interpretations derived from the analysis of results.

Furthermore, the structuring of the research questions needed to support the development of data preparation and processing techniques required to create systems and tools to facilitate the analysis of results and reporting data archived as hard copy documents with the EAU.

To this end the following sub-questions have been proposed to underpin the research and findings:

i. What framework of mechanisms, tools and procedures needs to be developed to aggregate, process, and analyse the available Junior Lyceum Entrance Examination (JLEE) data and reporting records?
ii. For the period 1997 – 2010 during which the NMC and FACTS policies were introduced,
    a. did the validity of the JLEE test constructs change over the years, in terms of continuity and consistency (Parallelism)?
    b. were there changes to the JLEE test forms over time that modified the complexity of these examinations and subsequently the mental load (Difficulty Level)?

iii. As a result of the introduced policies and considering the outcomes of sub-question (ii), has there been an impact on JLEE student outcomes that might reflect an overall improvement in the quality of education for these students?

4.2.2 Rationale

The focus of the study was therefore centred on policy impact on Year Six student attainment, specifically on the JLEE between 1997 to 2010. The selection of this set of criteria was based on the availability of, and accessibility to results and reporting data associated with this particular set of examinations. Additionally, the high-stakes nature of the JLEEs within the Maltese educational system made them a key yardstick that could underpin a longitudinal impact study.

The first sub-question is structured to prepare the available sets of data for analysis. Its purpose is to define the procedures used to process the data sets, organise them in a manner that is conducive to effective analysis and retain data quality and objectivity. Sub question (i) became a crucial part of the main RQ owing to the nature of the available data, which as stated earlier was only available as a printed document of results and reports.

In considering the associated second and third sub-questions, this study concurs with arguments by Creemers & Kyriakides (2007) in recognising that policy-driven change is a complex multidimensional construct. Using the JLEE results as an indicator of change is therefore only part of what would need to be a broader, more holistic investigation.

Following on from Creemers & Kyriakides (2007) discussion, however, large-scale policy design and implementation is purposed to have an impact on the quality of education as a whole. This is an argument also posited by Bezzina (2003) and Supovitz & Taylor (2005) and implies the possibility of detecting longitudinal variations in achievement on benchmark examinations as a result of changing educational quality, as long as the examinations retain construct validity over time. Sub-question (iii) is purposed to investigate those variations through an examination of student achievement scores in each of the five JLEE subjects for the fourteen years in question.

The construct validity over this same period, on the other hand, is addressed through sub-question (ii). Using benchmark examination outcomes to determine any impact on student learning requires that the study address factors deriving from intrinsic or extraneous influences from the examination sets themselves. This would render a clearer interpretation of any longitudinal comparative analysis of policy impact on outcomes. An analysis of the

medium of examination consequently preceded the analysis of results to consider such influencing factors.

The study therefore looked at variations in characteristics of the test constructs as the main intrinsic factors affecting construct validity. Any variations needed to be understood to determine if they were a result, intended or otherwise, of the changing policies and therefore imply affected JLEE processes and contexts. Similarly, analysis of the associated test forms reflected extraneous factors. This then became a key consideration as variations in the difficulty levels and discrimination power of the test papers would have a direct impact on the experienced mental load of the papers as a whole and subsequently on the outputs.

As discussed in the literature review, due to the longitudinal nature of the study, it was not feasible to consider the effect of germane factors associated with the different cohorts over the years.

4.2.3 Ethical considerations

This study was mainly based on an analysis of secondary data supplied by the EAU and did not deal directly with participants or schools, nor did the research require any form of personal or school information that could subsequently be linked to individuals. Nonetheless, consideration was given to determine possible associated ethical issues and data protection measures in accordance with Ethical Guidelines for Educational Research, Fourth Edition (2018) and the University of Malta research code of practice (University of Malta, n.d.).

Following approval to conduct the research by the MEDE research and innovation department, these ethical considerations were discussed with the director of the EAU. It was agreed that the scanned data provided by the EAU would be anonymised during the digitisation process conducted as part of this research. The outcomes of this process were a collection of disaggregated results that would be returned to the EAU once the study was over.

4.3 An integrated methodology

The nature of this study is concerned with understanding changes in student outcomes as a complex function of policy changes implemented over a fixed period. The context of the investigation established by the primary RQ requires both an understanding of the policy objectives and an associated measure of resulting change reflected in student outcomes. It is therefore necessary to determine if targeted intentions and defined improvement criteria established by the policies were in fact effective and, if so, determine to what extent. These qualitative and quantitative requirements, and their requisite overlap, suggest that an

integrated methods approach will respond to both simultaneously and highlight, in a clearer fashion, any relationships that may exist.

In considering the context of investigating policy influence on student outcomes, the combination of documentary analysis with an analysis of attainment data was determined to be a more relevant approach merging the methodologies used for process and outcomes studies. Work by Azorín & Cameron (2010) and Bartholomew & Brown (2012), considers integrating methods in such contexts to provide "*deeper exploration of causal mechanisms"* (Anderson, 2016, p. 236). Moreover, combining methodologies plays a key role in the study of education policies (Halpin, 1994, p. 205) and proves relevant to the analysis of policy implementation and impact while underpinning systems of accountability and data-driven policy development (Datnow & Park, 2010).

This research integrates an examination of literature and policy documents associated with the educational changes under consideration with an analysis of archival reports and outcomes data from student performance on an annual national examination. This choice of combined methodologies is derived from the research questions and consequently intended to establish a mechanism that combines these two sets of data by overlaying their time sequences for comparative purposes. By adopting an integrated methodology, the study sets out to establish context and intent through the analysis of the policy documents and from that analysis determine implementation timelines. This information is then compared to a time series representation of student achievement results to determine if any variation in those results follows policy implementation points along those same timelines. Furthermore, a comparative analysis of a third set of data from the EAU reports is used to inform context on variations to test constructs and forms thus rendering a comparative historical view of events and effects.

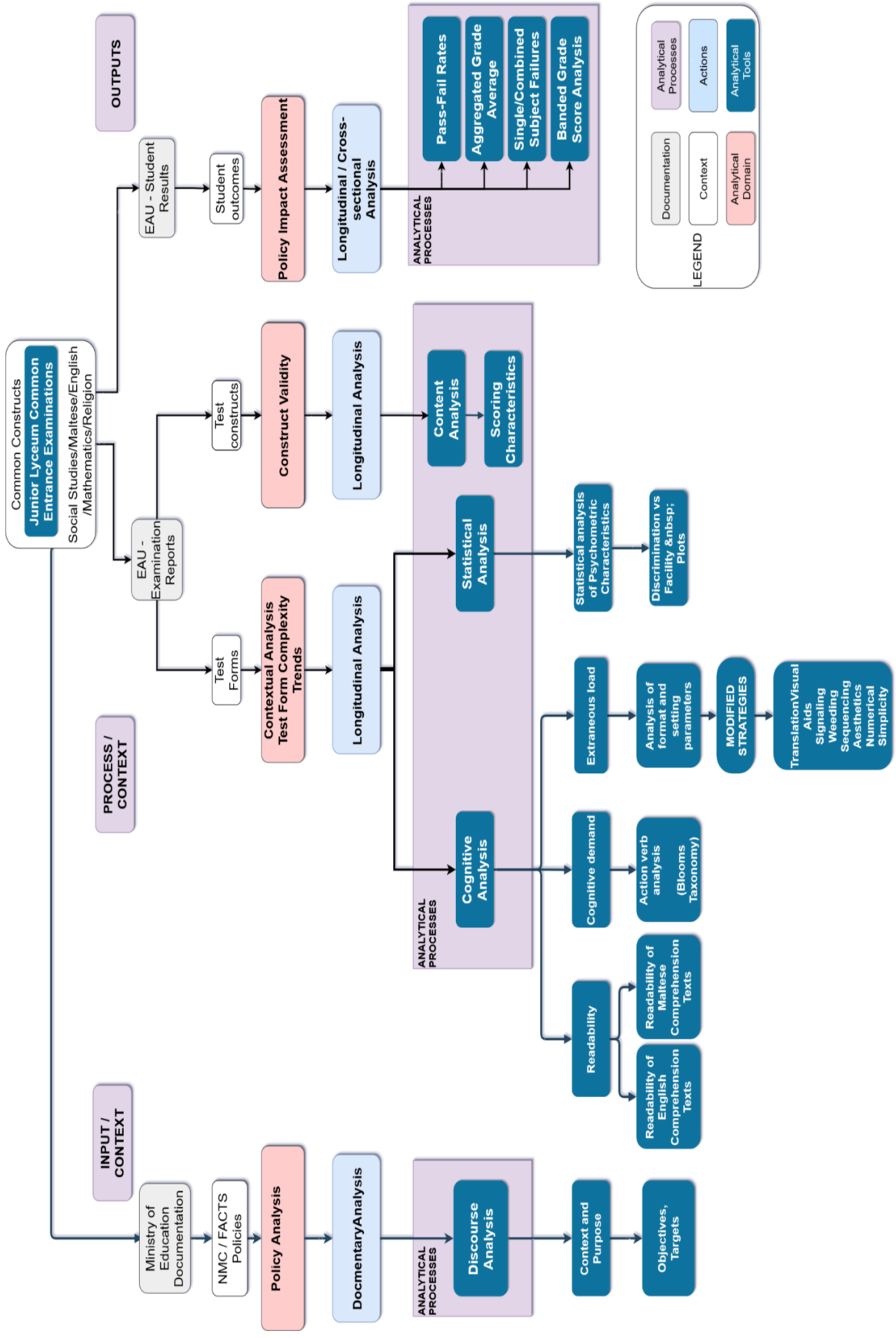
However, this study has certain limitations. Notably, it is only able to observe sequential trends and does not allow for direct causal associations between different strands of large-scale policy implementation and any particular variations in outcomes. The exact set of causal factors that lead to any change in outcomes is, as argued by Creemers & Kyriakides (2007), a result of the multidimensional set of influencing factors, some of which may be external or unintended and not necessarily related to the policy. Moreover, considering that the data being collected ranges from 1997 to 2010, the degree of external stimuli deriving from societal changes during different periods and affecting germane student characteristics cannot be determined effectively as part of this retrospective study. This latter limitation to the research

runs along similar lines as those argued by Newton (1997) and Patrick (1996), that meaningful longitudinal comparisons lose validity over the longer-term.

Nevertheless, the analysis will be able to shed light on longitudinal variations in the trend patterns of outcomes for the five different subject-based exams and compare those timelines to the policy introduction points while also comparing changes to test constructs and forms on the same time series. The integration of these various methodological systems of analysis requires multiple analytical streams running in parallel. These pathways are structured according to the research framework and organised into the three main domains underpinning the study: Input/Context; Process/Context; Outputs. The first focuses specifically on the policy documents, the second presents an analysis of the associated reports and the last stream is an analysis of outcomes and results. As this is a retroactive study of documented evidence, all three streams are initially subjected to a documentary analysis. The EAU reports and records of results are subsequently subjected to statistical analysis and all three can then be compared as discussed above.

The different analytical pathways are presented in a schematic flow diagram (Figure 4-1) below before being considered in further detail in the following sections.

Figure 4-1 Schematic of the overall analytical processes



4.4 Documentary analysis – Policies, reports, and records

This section discusses the analytical systems associated with each set of documents and elaborates on those methods applied to the policy documents and those used to analyse the yearly (EAU) examination reports and record of results.

There were three types of documentation considered relevant to this research. Each had a different function and served a different purpose, consequently needing a different type of analysis. The first was an analysis of the two main policy documents (MEYE, 1999, 2004a) that established and articulated the introduced policies, explaining all related aspects and issues. The latest educational policy change (MEDE, 2012), although not a key resource underpinning this research, was similarly scrutinised for comparative reasons that might shed further insights on common thematic threads across the years.

The second set of documents was an associated examination report (Department of Curriculum Development, Implementation and Review & Educational Assessment Unit, 1997-2010) for each examination cycle that included macro-level statistical data and subject-based examination reports. These played a key role in informing the contextual analysis as it related to the quality framework by working to understand variations in quality and standards of the examination sets.

The third area of analysis pertained to the documented record of examination results (Educational Assessment Unit, 1997–2010) which consisted of the actual record of outcomes and results for each student in five different subjects — social studies, Maltese, English, mathematics, and religion. This constituted the primary set of achievement data that was to inform the time series analysis. Subsequently, the document analysis for these records required different consideration and processing procedures from that of the policies and one that related to the use of these documents as a source of statistical data.

Moreover, these documents needed to be digitised before they could be processed and integrated for analytical purposes. This record of results for students undertaking the transitional examinations from Year 6 to Year 7, was not available as digital softcopies but was kept in printed hard copy form, bound, and stored. Similarly, the examination reports were available in printed form and the item analysis also needed to be digitised for further statistical analysis. These digitisation procedures are also linked to the first research sub-question considering methods for aggregating and processing the available data record and are duly considered and elaborated on in the last chapter of this section (Digitisation Process).

4.4.1 Policy documents – Establishing context and purpose

The NMC and FACTS policies were designed and intended to bring about paradigm shifts to the entire educational system in Malta and Gozo and modernise pedagogical practices, bringing a change to systems of teaching and learning (Calleja & Grima, 2012; Cristina, 2012; Galea, 1999, 2004; MEYE, 1999, 2004a; Mizzi, 1999, 2004). The analysis of these policy documents was therefore necessary to interpret purpose and meaning, determine set targets and understand the intended effect on schools, teachers, and students. The third policy prepared in 2012 (MEDE, 2012), although not an integral part of this analysis, would reflect on organisational intention through possible thematic continuity.

Anderson & Holloway (2018) have argued that an analysis of education policy *"lends itself to discursive exploration… by how it is constituted… and how it is taken up".* This supports the centrality of the two major policy documents to the purpose of this study as incorporated in the RQs. As this work is concerned particularly with the effect of institutional change on student outcomes, an analysis of the policies was used to determine the associated scope and intentions thus providing what Bowen (2009, p. 5) describes as a comparative context. Bowen (2009, p. 27) also posits that document analysis can be used to draw empirical knowledge and any systematic evaluation will establish implementation timelines for any defined targets, benchmarks, or indicators against which to measure success and effect. This analysis can therefore serve as a backdrop against which to articulate the level of policy implementation and determine effects.

Similarly, Taylor et al. 1997 (p. 37) argue that critical policy analysis is an effective way to establish any links to associated change and reform processes and determine *"whether and in what ways policies help make things better"* (Henry, 1993, p. 104; Taylor et al. 1997a, p. 37). Taylor et al.'s (1997) consideration of policy analysis proposes a three-level framework of analysis associated with understanding policies in general and extracting useful information from such documents - *"contexts, texts and consequences"* (Taylor et al. 1997a, p. 44). This framework allows consideration of these three domains to establish answers associated with the "Why?", "How?" and "To what effect?" respectively. The first and last of these questions underpinned the policy analysis, articulating an understanding of meaning of content and intended outputs and purposed to shed light on the intentions of those who fashioned them, thus allowing a clearer understanding of their goals (Codd, 1988; Hill & Varone, 2016, p. 5).

Subsequently, in conducting the analysis, this study focused on the content and outputs as they are integrated into the general discourse of the policy. Codd (1988, p. 243) discusses the

"implicit patterns" within the policies and holds that in order to start to deconstruct such documents, there needs to be an *"explicit recognition of the context"* (1988, p. 244). So, rather than looking at the nuts and bolts expressed on an interactional or institutional level, the analysis focused on the general discourse outlining direction, intention and goals that shaped the purpose and informed the desired "new" direction, what Anderson & Holloway (2018, p. 2) refer to as *"macro-social phenomenon"*.

In trying to establish a contextual backdrop, the policy analysis therefore centred on:

i. Briefly explaining the new educational context established by the policies.
ii. Identifying the purpose and objectives stated in each of the policies.
iii. Determining targets set to define success criteria against which to determine impact.

4.4.2 Examination reports - Specification grids and item analysis

The examination reports issued following each examination cycle gave a complete overview of that cycle outlining principally: general information; demographics and eligible population; statistical summations; historical statistics (dating back to 1981); rules and regulations; performance statistics; item analysis; specification grids; a copy of each test form; marking schemes; and summary reports from the chief examiner (Department of Curriculum Development, Implementation and Review & Educational Assessment Unit, 1997 - 2010). These reports were analysed to determine the degree of construct continuity over the years and determine longitudinal subject-based threads. They also proved valuable in reviewing variations in test form complexity and difficulty levels affecting mental load for each examination in the series. This in turn would form part of an analysis to identify longitudinal variations in assessment quality over the fourteen years.

In working to determine the level of longitudinal continuity and consistency of the test constructs, the study reviewed the information documented in the specification grids and item analysis for each of the five examinations. The specification grids were developed by paper setters before setting each test to *"describe the achievement domain being measured and provide guidelines for obtaining a representative sample of test tasks"* (Educational Assessment Unit, n.d.). These specification grids and the associated marking schemes (Educational Assessment Unit, 1997 - 2010) proved important for comparative considerations of the test constructs and forms over the period in question as they retained the same, or similar, format from 1998 till 2010. More specifically, a review of the grids could be conducted to understand variations in planned content and scoring distributions and support a

longitudinal comparison of the exams, identifying construct or form differences for each subject across the years.

The item analysis, on the other hand, was initiated in 1999 and continued through 2010, reporting facility and discrimination indices — with a similar purpose to those described by Aiken (1979) and Matlock-Hetzel (1997) — for each question of each exam. These were calculated from a random sample of 200 papers and offered a time series dataset that could be compared longitudinally to support the analysis of context and achievement.

The item analysis tables, however, were not available in digital form and a procedure similar to the digitising of results (discussed below) was applied to convert the data for these indices into a digital format. In all, twelve Excel files were produced—one for each cycle—that contained the facility and discrimination index for each question of the examinations ordered in the associated pair of indices. Once saved the data could be analysed as a set of time series for facility and discrimination.

4.4.3 Record of results – Preparation for analysis

The third set of documents was the collection of examination results achieved by each student who sat for the transitional examination over the fourteen years (1997 – 2010). This data presented the key resource for evaluating longitudinal changes in student outcomes. The scores were recorded and documented with the Ministry of Education's Assessment Unit in printed form and required a request be made to the Ministry of Education and Employment (MEDE) for permission to access, digitise and analyse the data in question.

Each yearly set of outcomes data was stored on two separate bound volumes each of which needed to be converted to a single-digitised file. Throughout this research, the original printed records are referred to as 'source records' while the term 'digitised records' will refer to those records resulting from the process of digitisation (Commonwealth of Australia, 2013, p. 5).

The methods used to digitise and prepare the 14-year record of achievement to respond to part of the research sub-question (i) presented earlier are described in a more structured manner in the Chapter 6. Once the records and item analysis were digitised, the data could be processed and analysed as a time series to identify fluctuations in the records and trend variations in the item analysis patterns.

4.5 The Data

The different data used in this study are described, defined, and considered in terms of applicability, validity, and reliability. As this is a retrospective study, all records and documents

were drawn from existing government sources and as such classified as secondary data sources (Hakim, 1982; Logan, 2020; Panchenko & Samovilova, 2020; E. Smith, 2008) except the policies themselves that are considered a primary source (Gall et al. 1996). These archived records and reports data are made up of student achievement scores collected during the years preceding, and those following, the introduced policy changes together with their associated examination reports.

For all items of record and documentation, Scott's (1990, p.19) criteria of authenticity, credibility, representativeness, and meaning were applied to the documents in question to test their validity and reliability.

4.5.1 Selection of JLEE data

As the scope of this research is intended to compare the impact of policy on student outcomes before and after the introduction of the policies, the design needed to investigate student achievement taken from a continuous series of standardised-test results. The Junior Lyceum Entrance Examinations record proved to be more consistent, dependable, and accessible when compared to other sources of data.

In deciding the best data sources to support this research, different datasets were initially considered to determine which of them offered a continuous run of achievement results over a prolonged period. The target cohorts were students enrolled in either primary or secondary state schools, or a wider cache of students from both systems. The data also needed to have two key characteristics to support the study: they offered a repeat cross-sectional window (Vignoles & Dex, 2007, p. 260) that would support a time series comparison of achievement; and they spanned the NMC and FACTS policies.

Secondary matriculation exams (MATSEC) and formal end-of-year school examinations were initially considered, however, they carried certain obstacles in terms of accessibility (for matriculation exams) and continuity and reliability (for school-based exams). On the other hand, the record of results from the JLEE across fourteen years to 2010 presented a set of data that was accessible, effectively documented, and had a consistent and continuous set of reports that explained the context from one year to the next. Furthermore, the record of achievement results, although not in soft copy, was considered to be a complete record over the years.

The selection of the JLEE data sets was therefore based on an initial perception of reliability of both the data and supporting documents to underpin a longitudinal study that could identify variations in achievement. That perception was corroborated by a deeper consideration of the

record of results that confirmed the JLEE data offered a facility to analyse a long run (14 years) of student scores, for five different subjects, which were collected in a regular and routine manner. The availability of this long run of methodically collected results offered an ideal source of evidence that could be applied to time series analysis for determining long-term impact (Glass, 1997; Gray et al. 2003; Lagarde, 2012) and be overlayed with similar time analysis of the policy introductions.

There were limitations on the research as a result of using one set of data however, namely that any conclusions could only effectively relate to the Junior school population and could not be linked to secondary or post-secondary populations without some form of correlational investigation.

4.5.1.1 Sample size

As Malta has a relatively small student population that is measured in the tens of thousands (primary and secondary), it is possible to establish year group sets of a few thousand making up the entire population for that year group.

This outcomes data set included results for those students having undertaken the JLEE. The records available tabulated the scores of participants from all the state schools and those private and church schools who opted to participate. On average, 89% of Year 6 state school students, and 36.3% of the non-state school population applied to sit for the JLEE over the fourteen years being considered (Table 9-1 Year VI student population and cohort numbers by year). This constituted a total of 72% of the entire Year 6 population across the islands and can be considered to be a representative population sample. This would also hold true for the associated annual reports issued by the EAU.

In considering the non-applicants, this research did not investigate the reasons for students not applying to sit the JLEE. However, it is noted that most non-state school students would have had the option to continue their education by moving directly from primary to secondary within the same school without any need of sitting for an examination. Most of the non-applicants from state schools would have chosen to continue to state area schools rather than sit the JLEE. The reasons for these choices are unspecified in the literature and beyond the scope of this research.

4.5.1.2 Characteristics of available data sources

All sources of data and documentation used in this study were accessed through official government sources in the form of printed hard copies or digital soft copies.

i. Policy Documents (Soft copy) –- first accessed online (05/09/2016) https://education.gov.mt/en/resources/Pages/Policy-Documentation-Archive.aspx
ii. Examination reports (hardcopies) — 14 Publications (Curriculum Department & Educational Assessment Unit, 1997-2010) — accessed through the Educational Assessment Unit
iii. Record of results (hardcopies) — 14 sets — accessed from the Educational Assessment Unit

As stated earlier, the first set of (policy) documents was intended to set the scope of the changes while the analysis of achievement and quality required the use of data recorded in the latter two sets. The reports offered documentary evidence describing the exam context and year-on-year statistics while the record of results provided a more granular set of data presenting micro-level data on individual scores for each of the examinations.

Although the policy documents are considered primary sources, both the reports and the record of results are secondary data sources and applicable to longitudinal observation (Cave & von Stumm, 2021; Siddiqui, 2019). Vignoles & Dex (2007) have discussed the suitability of secondary sources of both micro- and macro-level statistics for time series analysis and proposed various applications associated with understanding longitudinal and cross-sectional change. In the context of this study, the applied usage is a hybrid version they refer to as *"repeat cross-section data"* (2007, p. 260) using student achievement data for the main analysis of outputs and the reports for understanding possible changes in process and context.

4.5.2 Suitability of the data

4.5.2.1 Examination reports

Each of the specification grids for the JLEE was at hand from 1998 – 2010, as were the test forms with their associated marking schemes. The item analysis for each examination was available from 1999 – 2010. This information was all documented in the EAU's published annual reports (Curriculum Department & Educational Assessment Unit, 1997 - 2010) and thus considered authentic, credible and representative (Scott, 1990).

4.5.2.2 Record of results

The sets of achievement data used in this study were drawn from records of end of (Year Six) Primary School transitional exam results (1997-2010). The record of results was compiled as part of the EAU's remit to maintain a formal government record of results for all students

sitting for the JLEE examinations. As such it was also considered authentic, credible and representative (Scott, 1990).

4.5.2.3 Validity and reliability

The data being used is drawn from the official record and reports published by the educational authorities relating to the yearly examinations under consideration. They cover the majority of the Year 6 student population in government schools and offer a complete record of results between 1997 and 2010 (and further back for context), for the different subjects.

As the record is drawn from official government documents used to formally report on centrally-controlled assessments, they are considered highly reliable (Siddiqui, 2019) factual sources that can be purposed for longitudinal, comparative analysis (Goldstein, 2001, p. 434).

The available data being proposed here has been meticulously collected and recorded through formal institutional processes and procedures. Furthermore, the sample size being studied is a substantial proportion of the Year 6 state school cohort (approximately 89%). Both these factors strengthen the validity of the interpretation of variations in outcomes related to policy change (Kreber & Brook, 2010, p. 99).

4.6 Summary

Aggregated, the general design established three analytical pathways that needed to be followed to render a more detailed analysis underpinned by the quality framework. This design was established on the premise that the research required both a documentary analysis of the policies, a data analysis of the record of outcomes and a combination of documentary and data analytical procedures for the EAU examination reports (Figure 4-1). The selection of the JLEE data was based on the characteristics of the examinations themselves that enabled the analytical design and were straightforward to access. Furthermore, they offered a complete collection of benchmark examinations that spanned the two key policies and as such could support a longitudinal impact study.

5 Methodology

5.1 Chapter overview

Following on earlier discussions, although the test forms may have been developed according to the same content and statistical specifications by the same developers, this did not guarantee a consistent level of difficulty for successive test forms. As postulated by Alberts (2001); Coe (2010) and Kolen & Brennan (2014), it remains highly likely that the true difficulty level of the test form with the same specifications varied despite efforts to maintain similarity through the test form design. It has also been suggested by Coe et al. (2008), Dorans et al. (2010) and Fitz-Gibbon & Vincent (1997) that the likelihood of categorically determining the actual difficulty level of a test form is not possible in any conclusive way. However, what has been argued earlier on is that the application of cognitive load theory could be used to shed light on trend variations in subsequent test forms that would give more depth to any analysis of policy impact on outcomes.

The proposed analysis endeavours to determine variations in intrinsic and extraneous cognitive load characteristics of each of the test forms to determine longitudinal trend variations that could then be considered against the respective analysis of outcomes.

The first section of this chapter considers the practicality of time series as the ideal mechanism to underpin a longitudinal impact analysis. It presents arguments, drawn from the literature, for using these methods in determining educational effectiveness associated with changes in educational systems and associates its relevance to the context of the study.

The second section is structured to discuss the methodologies used to analyse the validity of the test construct over the period in question and the trend analysis of variations in test form complexity. This latter analysis became essential to the study in order to understand

longitudinal variations in test form difficulty levels and their effect on student outcomes and required a combination of comparative trend analysis of several factors affecting the mental load exerted by the JLEE test forms.

Section three establishes the pretext for using achievement results to determine longitudinal change in the quality of education before outlining the methodological processes used to analyse the digitised records of the JLEE results. It also defines the period for analysis and reviews the application of time-series analysis to understand variations in student achievement over that time.

5.2 Analytical design

Considering the characteristics of the data being used and the research questions being asked, this section proposes an analytical design that requires the application of longitudinal methodologies. Longitudinal research methods and time-series analysis lend themselves to understanding change over time and will allow an understanding of the rates at which those changes take place (Box-Steffensmeier et al. 2014; Collins, 2006; Feldhoff & Radisch, 2021; Singer et al. 2003; Wei, 2006; Wu et al. 2013).

From the data gathered, the analysis intended to establish a structured investigative model underpinned by statistical mechanisms that identified variations in student achievement over time. The measuring of such changes is fairly straightforward as the associated longitudinal data is readily available (Singer et al. 2003). Similar assertions are made for studying change using time series analysis by Box-Steffensmeier et al. (2014) and Glass (2006). Similar to arguments posited by Wei (2006, p. 458), the study will establish a *"time-ordered sequence of observations"* and generate a timeline populated with fixed point events. Outcomes data can then be analysed for sequential patterns of association rendering possible insights into correlational, if not causal, relationships (Salkind, 2010, p. 1521). As the time of the specified intervention – the introduction of a new national policy – is known, then any resulting impact of the intervention will be reflected in a time series by a change in intercept or gradient of series at or around the time when that change took place (Shadish et al. 2002).

The importance of using longitudinal analysis to get an overview of the impact on educational effectiveness resulting from changes to the educational system is identified by Creemers & Kyriakides (2007) as one of two methodological imperatives. Such analysis is predicated on the research purpose and questions (Azorín & Cameron, 2010; Morrison, 2007; Teddlie & Tashakkori, 2009) and offers an effective approach to responding to those questions. Applied

to the context of this study, such an approach is centred on two singular fixed-point events established at each of the policy launch dates.

However, there is an associated premise that the implementation of these policies established a context for change that was intended to create positive pressure leading to a general improvement in the quality of education and learning as declared in the policy documents themselves (MEYE, 1999, 2004a). Bezzina (2003) and Supovitz & Taylor (2005) have argued that such policies act to improve the overall quality of education systems that in turn improve student outputs and subsequently outcomes. Similarly, Adams (1993), UNESCO (2005) and the World Bank (n.d.) have also acknowledged that general improvement of academic standards can in part be reflected in improved learning outcomes and linked to an improvement in the quality of learning and education. These arguments do not imply an absolute association whereby improved quality is necessarily reflected in improved achievement. Rather, what is being suggested here is that if the quality of education and learning does increase within an educational context, then there is a strong likelihood that achievement will follow suit and subsequently be reflected in overall student outcomes.

Nonetheless, the ideal analytical design for investigating any impact remains a longitudinal study. This would principally take the form of an overlapping time series study of inputs and outputs, further informed by parallel time series analysis of process and context.

5.2.1 Longitudinal data analysis

The literature review presented arguments in favour of using longitudinal analysis to get an overview of variations in educational effectiveness brought about by changes to the educational system. It is however the nature of the available data sets that give greater weight to this choice of methods.

As stated in the previous section, Creemers & Kyriakides, (2007, p. 5) argued that longitudinal analysis is considered one of two methodological imperatives in such situations. Their second imperative —multilevel organisational structure analysis—is concerned with macro- and micro-level analysis taking place within the schools. As this research concerns itself with the broader overall effect resulting from national policy change on student outcomes, the application of multilevel analysis was considered beyond the purpose of the study.

Furthermore, White (2010) and Yin, 2006 (p. 43) have argued that although the selection of research and analysis tools for impact evaluation is context dependent when selecting particular processes, most impact studies tend to follow quantitative approaches which *"… are*

often the most appropriate methods for evaluating the impact of a large range of interventions" (White, 2010, p. 153).

More specific to the context of this research, Glass (1997, p. 4) points out that time series is an effective way to process and analyse achievement data to determine the effects of an implemented change or treatment. He clarifies that in these situations the data collected is not purposed to serve an experimental framework but can be used to inform analysis. Siddiqui's (2019) discussion of the pros and cons of using such secondary data in research, also supports these points of view arguing that longitudinal datasets are invaluable sources that can be drawn on to render patterns and trends and identify fluctuations or variations over time.

These arguments, together with those presented in the literature review, all support the longitudinal analytical design being proposed to investigate the impact of the NMC and FACTS policy on student outcomes.

5.2.2 Time series analysis

In analysing archival results and officially reported data sets collected over 14 years from 1997 – 2010, this study is therefore exploring policy impact on student outcomes. More specifically, the data under investigation was associated with a nationally administered qualifying set of examinations (Grima et al. 2008, p. 29). The nature of such exams made it inappropriate to apply any controlled experimentation procedures at the time and considering that this research is an analysis of past records such procedures are not an option.

The arguments from the previous sections have suggested that the more practical method of study to serve the main research purpose would have a time series design. Work by Biglan et al. (2000), Kontopantelis et al. (2015) and Teddlie & Tashakkori (2009) support such arguments proposing a quasi-experimental design with a time series analysis of associated results data before and after the fixed-point events. Arguments by Biglan et al. (2000), Glass (1997) and Kontopantelis et al. (2015) support the choices being made here arguing that such methods can be applied to this type of data to determine short-term impact and long-term trend differences, linking input changes with output variations.

However, in the context of this study, the expectation was not to observe an immediate jump in outcomes, but a gradual or time-lagged change in the number of successful candidates undertaking the examinations over time. This expectation was due to the large-scale nature of the policies and the inertia (Thomas, 2002) associated with the implementation of effective change that would manifest as lagging trend variations.

In determining the analytical design, initial consideration was given to an interrupted time series approach. Wagner, Soumerai, Zhang, & Ross-Degnan (2002, p. 299) and Glass (1997, p. 1) have argued that interrupted time series methods offer one of the stronger quasi-experimental methods that can be used for longitudinal research on par with more randomised experimental methods. However, according to Shadish et al. (2006, p. 546), interrupted time series tend to be applied to interventions or treatments that are distinct and unitary and this was not the case for the introduction of both the NMC and FACTS policies. Although both policies were introduced at a specific point in time, they were intended to be multi-variate in their implementation and "holistic" in their goals as described in their introductory texts (MEYE, 1999, 2004a).

The complexity of introducing these long-term policies meant that the implementation of each would be a sequence of intermittent changes with some being introduced immediately and others over a longer period. Furthermore, such policy introductions would take place as sequences of associated or disjointed initiatives rather than a singular intervention. They would also be subject to feedback and respective adjustment actions during their implementation phase. These factors meant that no one change resulting from the policies could be identified as a distinct intervention but rather, each policy needed to be considered a singular, multidimensional treatment.

Time series analysis, therefore, played a key role in the analytical processing of the datasets and was applied to view variations in the student achievement results (Pass-Fail rates), as well as changes in test constructs and test form complexity, across the two fixed points defined by the implementation year of each policy. By analysing data collected at the time and comparing it to the period before and after the policy introduction, the study was able to identify longitudinal year-on-year trend variations.

Furthermore, together with the structure of the datasets, the analytical design supported cross-sectional comparisons of long-term trend variations across the five different subjects and was used to deliver comparative insights into subject-specific trends relative to one another. This cross-sectional analysis was also established around the fixed-point markers defined by the policy introductions and allowed a subject-based impact comparison.

5.2.3 Limitations and considerations

The analytical structures being proposed in this chapter establish a quality framework of reference, similar to that described by UNESCO (2002, 2005), that lends itself to investigating change in terms of input, process, output and context. This can be applied to interpret

characteristics of quality in educational contexts but has its limitations (Scheerens et al. 2011a).

The key variables analysed — student outcomes represented by the pass/fail rates, psychometric variations, and changes in mental loads — are, according to Box-Steffensmeier et al. (2014), contextually dependent on other social processes and are influenced by a *"temporal dependency between social processes"* (2014, p. 8). This argument is similar to that made by Newton (1997) and Patrick (1996) that comparative meaning is less valid over longer periods due to changing contexts and cultures. As such, the variables are associated with a multiplicity of factors that influence student outcomes as different cohorts progress through the same learning environment and as a result, changes in outcomes need to be interpreted accordingly. These complexities are discussed by Feldhoff & Radisch (2021) in association with school improvement research and underscore the fact that any linkages determined by this research remain strictly between policy and outcomes on the JLEE. Similar arguments by Collins (2006), imply that the nature of such studies does not allow for direct causal associations to be determined with absolute certainty.

However, national policy changes like the NMC and FACTS are overarching in scope and designed to be effective over a longer time. They tend to target and influence, directly or indirectly, the various contexts impacting the multiplicity of factors and are complex in nature (Feldhoff & Radisch, 2021; Stevenson, 2003, p. 11). Although such macro policies may have an immediate effect in the short-term, they are usually designed to impact the educational landscape as a whole. Codd (1988) and Hill & Varone (2016), argued that the broad-scale nature of such policies means that they tend to be designed with longer-term goals and implementation strategies in mind. This would suggest that although analysis of learning outcomes may be more meaningful for shorter-term impact, long-term trend variations may reflect the wider overall influences on those contexts and cultures.

Reports by Said (2015) on the NMC and Borg & Giordmaina (2012) about FACTS confirm the intention to improve Maltese education over the long-term (see 2.3.4 Systemic reform process in Malta), mindfully considering a forecast of needs and changing cultures. The case for these two policies is reflective of long-term planning, preparation and implementation and it follows that the different influences resulting from the changes work, to varying degrees, to deliver a positive influence on the whole learning environment. These intentions are stipulated by the responsible Minister in each of the policy documents (MEYE, 1999, p. 2,3, 2004a, pp. xi–xiii).

As such, the broad-scale policy documents were prepared to render a positive overall influence on the quality of education and learning and it can therefore be hypothesised that the implementation created positive pressure on the system as a whole. The research is therefore limited to considering the broad-scale impact of the policies and investigating trend variations in process, context and outcomes associated with the JLEE alone.

5.2.4 Analytical framework

This section presented arguments for an analytical design that would respond to the RQs and available data sets. The proposed methodology for this study assimilates a structure similar to the one used by Collins (2006) in that the policy analysis (discussed in 4.4.1 above) establishes the context describing the intended change by fixing the implementation dates on a timeline; the design is a time series analysis of outcomes and; the analysis attempts to render a statistical understanding of impact and effect on those outcomes.

In the context of the quality framework (i-p-o-c), an analysis of outputs will also reflect on the overall variation in quality as long as the systems of assessment maintain continuity and consistency. This latter consideration becomes an important influencing factor as it recognises that stronger similarity between constructs implies closer association and therefore a better foundation for comparison (Coe, 2010; Kolen & Brennan, 2014; Newton, 2005). A longitudinal analysis of variation in test constructs and forms was subsequently conducted to understand if changing intrinsic or extraneous factors may have influenced the cognitive load of the examinations and consequently affected outcomes by altering the mental load (difficulty) of the questions This analysis was then used to support the interpretations drawn from the analytical framework that was established around the documentary analysis of the policies and the longitudinal data analysis of prior and subsequent result outcomes.

5.3 Longitudinal analysis of context: Test constructs and test forms

The review of the literature regarding the Junior Lyceum Examinations established that these sittings were developed and prepared by a central authority (MEYE, 1999, 2004a) as a qualifying national examination (Grima et al. 2008). Furthermore, each of the five examinations was intended to be structured around a common construct that may or may not have varied over the 14 years in question. Such a change would have implied a change in the mental load of the examinations affecting complexity, and subsequently difficulty levels. Therefore, in trying to understand the degree of parallelism between sittings, the study reviewed trend variations in both the test constructs and the test forms between 1997 and 2010.

Paas & Van Merriënboer (1994), Sweller et al. (1998) and DeLeeuw & Mayer (2008) have argued that the mental load is the combined effect of intrinsic and extraneous cognitive loads exerted on a learner and is associated with the complexity of a task. Consequently, in order to get a clear understanding of any change in complexity of the JLEE, it became necessary to focus on a longitudinal comparison of mental loads. This required comparing both the intrinsic and extraneous CLs over time.

The intrinsic CLs associated with the examination have been linked by DeLeeuw & Mayer (2008) and Kettler et al. (2009) to variations in the test constructs. On the other hand, extraneous CL has been associated with variations in examination format and design (Crisp & Novaković, 2009; Gillmor et al. 2015; Sweller, 1988). Both CLs affect the mental load of the test forms.

The literature review discussed a dual methodology suggested by Newman et al. (1988) for gathering insights into such trends. Their work proposed both cognitive and statistical methods to determine the difficulty levels of test papers. This study applied both these comparative mechanisms to investigate longitudinal trends in difficulty levels of subsequent test forms. The cognitive analysis applied cognitive load theory systems for processing and analysing extraneous factors manifest in the test papers. The statistical comparative made direct use of the facility and discrimination indices reported on each of the examination reports (Curriculum Department & Educational Assessment Unit, 1999 - 2010).

This section of the methodology therefore investigates mental loads through an analysis of continuity and consistency on the exams by investigating successive test constructs and associated test form structures. The purpose centred around identifying any longitudinal variations in the two key factors impacting mental loads as identified by DeLeeuw & Mayer (2008), Paas & Van Merriënboer (1994) and Sweller et al. (1998).

5.3.1 Linking constructs and comparing forms

This section looks at the two factors affecting mental load in more detail as they link to constructs and forms and explains why germane CL was not considered as part of the study.

This research considers the complexity and difficulty levels exerted on the test takers as a function of the intrinsic and extraneous CLs of the respective test constructs and test forms (3.4.3 above). In selecting these factors on which to focus, two considerations were made: firstly, as stated earlier, what data was readily available and accessible for analysis in the documentation and stored records; secondly would the available data inform the analysis as intended to determine the degree of variation over the years.

Factors associated with germane CL were not investigated. Germane CL associated with the students as individuals or groups — their prior knowledge, motivation, and approach to learning — is relevant to inform the teacher-student interaction as part of lesson development (DeLeeuw & Mayer, 2008; Sweller et al. 1998). It is not, however, as relevant to understanding trends reflected in the analysis of large student cohorts as represented in this study. Sweller et al.'s, (2019) more recent review of germane CL does not change this point of view. Additionally, the retrospective nature of the study makes it difficult to deliver any sort of analysis of the effects associated with germane CLs.

More relevant to the analysis being proposed, a comparison of results remains dependent on successive tests having comparable construct validity (Chapelle, 1998; Coe, 2010; Downing, 2003; Kane, 2013). Any variations in these constructs would need to be understood in the context of the introduced policies to determine if they were a result, intended or otherwise, of the policy changes. Establishing the degree of similarity of successive constructs, therefore, became a necessary first step in this process – a linking exercise — that would underscore further analysis and interpretation (Coe, 2010; Kane, 2013). This part of the analysis required that the five different test constructs administered to subsequent cohorts were first compared to determine longitudinal consistency and establish grounds for further comparisons.

Furthermore, as the discussion has suggested, an analysis of intrinsic CL can be drawn from the same longitudinal analysis of the test constructs. Any changes in the test construct would imply variation to the cognitive demands and require consideration of any impact on the overall mental load as a result of any changes made to test items on parallel forms. As complexity is a function of the *"number of elements that must be processed simultaneously"* (Sweller et al. 1998) an approximate measure of the complexity of a test construct was determined through an analysis of factors that define these interacting elements.

On the other hand, an analysis of the extraneous CL of the test forms was used to help understand variations in test form complexity. It was not possible to determine a definitive measure of the extraneous CL exerted by the test forms as that tends to have a subjective dimension with different individuals having different perspectives of difficulty. However, it was possible to establish comparative trends based on common objective baselines. This part of the analysis used a dual methodology to determine fluctuations: a cognitive analysis of the test forms and a comparative statistical analysis of the general psychometric characteristics reflected in the outcomes.

5.3.2 Construct continuity – Content and marking schemes analysis

Linking the constructs over the period in question was necessary to strengthen comparative reasoning and interpretation, and help determine the extent to which the test forms could be considered similar or parallel (Dorans et al. 2010; Feuer et al. 1998). Parallel test forms implied similar functions structured around the same or very similar constructs (Angoff, 1984; Dorans et al. 2007).

Coe (2010), Kolen & Brennan (2014) and Newton (2005) discuss situations where the test constructs are different and consider systems of linking test forms from different frameworks. The situation here, however, is more *"straightforward … designed to exactly the same framework and specifications"* (Newton, 2005, p. 107). The purpose of this research requires comparing construct validity arguments across the years to determine the degree of parallelism and consequently any variations in the intrinsic factors affecting the complexity of the test forms.

In a report on the assessment process used in the transition from primary to secondary education, Grima et al. (2008) asserted that the Junior Lyceum Examinations were prepared against a specification grid reflecting the *"knowledge skills and processes laid down by the respective primary school syllabi."* (2008, p. 94). These specification grids were included in the EAU annual reports and ensured that the distribution of marks was aligned with assessment criteria. As such, they established formal documentation that could be used to review construct validity and determine longitudinal parallelism in terms of continuity and consistency.

Furthermore, Grima et al. (2008, p. 107) noted that the English, Maltese and religion examinations undertaken in 2006 were based on new syllabi published in 2005, with mathematics changing in 2007 (Curriculum Department & Educational Assessment Unit, 2007, p. 51). These changes to the syllabi needed to be investigated to determine if the specification grids varied as a consequence, and if so to what degree. No record stipulating changes for social studies was found in the reports. Additionally, the 2010 sittings did not include social studies as part of the qualifying set of examinations.

The five annual exams were described by Grima et al. (2008) as high stakes, qualifying examinations and informally considered *"an important benchmark in our educational system"* (Curriculum Department & Educational Assessment Unit, 2005, p. v). They were structured to maintain continuity and preserve examination standards lending themselves to understanding the construct links and validity. To do so, however, required the structuring of a set of tools to

process the examinations, analyse the content and determine if they provided an "*internally consistent measure"* as discussed by Coe (2010, p. 279).

Considering that equating scores was not an option due to the absence of availability of raw scores, an analysis of construct validity needed to be structured around content analysis, statistical specifications and scoring characteristics. To better investigate if there were any variations in the test construct for each of the five different subjects —and subsequently, to the intrinsic CL of those constructs (DeLeeuw & Mayer, 2008; Gillmor et al. 2015; Kettler et al. 2009)— the study reviewed and compared:

i. The objectives and standards set by the examination boards and stipulated in each of the examination reports.
ii. Construct changes that may have been reflected in the scoring characteristics and anticipated difficulty levels.
iii. Curricular changes integrated into subsequent examination cycles.

5.3.2.1 Comparative content analysis

According to the American Educational Research Association et al. (2014, p. 15) a comparative content analysis across the years, can be leveraged to establish insights into variations in meaning or interpretation of achievement results across different student cohorts. In determining the "alignment" of test constructs to intended standards, a validity argument can be established for that construct and associated test forms (2014, p. 15).

In the context of this study, however, the alignment being investigated was not concerned with correspondence between the tests and the prescribed syllabi but looked at reliability and precision (continuity and consistency) in delivering the same test constructs over the years. In doing so the analysis attempted to determine the degree of parallelism or divergence reflected in the content structures by looking at specification grids and marking schemes reported in the EAU reports. Work by Bowen (2009) has argued that analysis of such documentation is an efficient way of providing context and tracking changes over time. The comparative tools were therefore structured to analyse the subject specification grids for the years 1998 - 2010 and identify any changes over that period. The 1997 reports did not have these specification grids and could not therefore be processed using these instruments.

Furthermore, in a detailed analysis of the JLEE, Grima et al. (2008) stated that all the preparation and structuring of the examinations up to 2007 were established on the state's primary school syllabi, implying a common thread linking subsequent construct definitions. There were, however, different publications of the respective syllabi that needed

consideration in the analysis to identify updates on the specification grids in general, and subsequently any effect on the constructs specifically.

In developing the specification grids, the EAU had issued guidelines (Educational Assessment Unit, n.d.) that detailed general procedures for paper setters stipulating that the grid would be used to indicate:

> *"(i) the learning outcomes to be tested.*
>
> *(ii) the subject matter or content area.*
>
> *(iii) the assigned weighting to the learning outcomes and content areas in terms of their relative importance"*
>
> *(Educational Assessment Unit, n.d., p. 3)*

The specification grids for each subject were required to list the subject matter and content areas drawn directly from the Year Six syllabus as well as the learning outcomes (objectives) being tested: *"(a) recall of knowledge (b) intellectual abilities or skills … (c) general skills … (d) attitudes, interests, appreciations"* (Educational Assessment Unit, n.d., p. 4). The grids also included mechanisms linking the question items or test sections to the respective syllabus and in most cases also included a difficulty level estimate for the test item or section.

Figure 5-1 below is an example of the specification grid used for social studies (Curriculum Department & Educational Assessment Unit, 1998) showing the main domains and subdomains. The referencing to the specific syllabus was, in the case of social studies, a separate table that linked the "Exam Paper Section" to the various parts of the syllabus.

Figure 5-1 Example of specification grid domains and sub-domains for Social Studies

Exam Paper Section	Objectives				Estimated Difficulty			Contributory Subjects		
	Knowledge	Understanding	Skills	Attitudes	Easy	Moderate	Difficult	Human Environment	Geographical Environment	Historical Environment
	Marks	Marks	Marks	Marks	Marks	Marks	Marks	Marks	Marks	Marks
A	7	1	2	-	10	-	-	-	-	10
B	6	1	2	1	10	-	-	-	10	-
C	6	1	2	1	10	-	-	10	-	-

This series of yearly specification grids therefore presented a continuous structure of domains and sub-domains, associated with learning outcomes and content areas, which could be used to support a comparative content analysis. As the details of each exam question or section were mapped onto the grid to define the construct characteristics according to a predefined blueprint, then any changes to that blueprint would show up on the specification grid for a particular year. Further analysis was initiated if and when a substantial redistribution of

weighting was identified across any of the defined sub-domains. This was done for the year before and following the change to understand if that change was an anomaly or constituted a more permanent change to the construct's statistical characteristics.

5.3.2.2 Scoring characteristics and anticipated difficulty

The scoring characteristics of each paper were integrated into the specification grids in different ways: English and mathematics had prepared granularly meticulous specification grids assigning marks down to a sub-question level; social studies and religion specification grids had allotted marks according to sectional grouping schemes; the Maltese examinations offered the least detail, simply check-marking the application of different learning outcomes without allocating score weighting. This required a slightly different approach to analysing the details for each of the subjects.

Furthermore, each of the subject reports, except for Maltese, had an anticipated difficulty level grid (Low, Moderate and High) that offered a statistical distribution of score weighting according to difficulty level. The anticipated level of difficulty provided another statistical specification that reflected on efforts of the test writers to moderate the difficulty levels of the test to suit a predetermined mix. In mathematics, they specifically recognised this effort in the reports (Curriculum Department & Educational Assessment Unit, 2005, p. 50), wanting to deliver a differentiated paper that had similar weighting distribution across all 13 years. The analysis of this distribution was used to identify if there were any changes in the planned distribution of difficulty level by paper.

However, even if a longitudinal analysis of content and scoring characteristics were to show a high degree of construct continuity and consistency, test difficulty still varies due to varying question styles or demands. Such variations in extraneous factors also impact the mental load and consequently the overall outcomes. The analysis therefore needed to include a deeper look at test form complexity to determine variations in extraneous cognitive loads. The two analytical methodologies mentioned earlier — cognitive and statistical — were subsequently developed to investigate such changes in complexity and are discussed in the next two sub-sections of this chapter.

5.3.3 Cognitive analysis — Determining trends in test form complexity

This section presents a methodological framework used for the cognitive analysis of the test forms affecting mental load, before discussing the statistical analysis in the following section.

After completing a comparative appraisal of subsequent test constructs, consideration then turned to understand the extraneous loads of the respective test forms. This study maintains that in determining relative trend variations in test form complexity, concepts of cognitive load theory can be applied to structure a comparative analysis of associated difficulty levels. CLT, although usually associated with instructional design, allows an investigation of changing difficulty levels to be underpinned by a theory that categorises various affecting factors and could be used to establish an analytical framework.

This part of the analysis was therefore intended to understand longitudinal trends in difficulty levels as a reflection of changing test form complexity rather than determine exact measures of complexity. The research draws on work by Brindley (1987), Candlin (1993), Nunan & Keobke (1995) and Sax et al. (1972) and maintains that for the different test forms underpinned by the same construct there exists a proportional relationship between task complexity and task difficulty level.

Based on the availability and nature of the data and original test forms the study established an analytical framework similar to that offered by Kettler et al. (2009) to analyse test form complexity. This was established around an analysis of readability levels, cognitive demands of test items, and the general format, structures, and presentation of the test forms. To elaborate, the cognitive analysis was structured around a comparison of:

i. Readability of the examination passages used for comprehension-type questions to determine changes in associated cognitive demands exerted by these passages (Gillmor et al. 2015, p. 4).
ii. The cognitive demand of the examinations as determined by the action verbs in each of the test form items (Dueñas et al. 2015; Jones et al. 2009).
iii. The general format and setting parameters for each of the subject examinations (Sweller et al. 1998, p. 263).

These three main areas could be compared longitudinally to determine variations. Any variation in the difficulty levels between test forms implied variation in complexity and was then considered to inform the impact study on student outcomes.

5.3.3.1 Variations on Readability scales

As this analysis is underpinned by a longitudinal comparison of test constructs and forms rather than an exact measure of their complexity then the application of a common readability scale to the test passages from subsequent test forms is enough to graphically determine variations. To this end, an analysis using readability algorithms was applied to determine

longitudinal changes in difficulty levels for Maltese and English comprehension text passages taken from the respective test forms.

This approach was established on the premise that there is a correlational link between readability and item difficulty on assessments (Gillmor et al. 2015; Hewitt & Homan, 2003). Other work Mifsud (2019) also applied readability scales to compare difficulty levels of comprehension texts in Maltese and English on international tests.

i. Readability of English texts

As there were no digital copies of the test forms available, each of the English comprehension texts on the exam papers (1997 – 2010) was scanned as a PDF and converted to an editable MS Word document. This was then analysed using various statistical tools and readability algorithms.

The analysis of comprehension texts taken from the English test forms was processed using a raft of six online readability algorithms hosted on the Text Readability Consensus Calculator (My Byline Media, 2020). This application delivered six grade-level readability scores using different formulae: Gunning Fog, Flesch-Kincaid Grade Level, The Coleman-Liau Index (CLI), The SMOG Index, Automated Readability Index (ARI), Linsear Write Formula. The ARI and CLI formulas are the only formulas that do not rely on syllabic features of the text in their calculations — making them independent of language characteristics — while the other algorithms do.

In looking for trend patterns over the 14 years (1997 – 2010) the analysis plotted the grade-level scale determined by each algorithm against the examination year (the grade level range ran from grade 2 to grade 12 rather than a percentage scale). All six plots were then collected on a common graph of grade level vs. year so that the outcomes for each of the algorithms could then be compared. This allowed the analysis to visualise trends and determine if the fluctuations in readability followed similar patterns when processed through the six different algorithms.

Furthermore, each grade level score was considered a unit data point derived from the respective readability tools and having a common outcome measure. This, according to Laird & Mosteller (1990) established a single measurement situation allowing the study to synthesise the six scores and determine a combined grade level average. Such a synthesis can be applied to multiple methods analysis to compensate for inherent bias (Levenson et al. 2000). Those averaged results were then plotted against the year of the sitting and analysed accordingly.

ii. Readability of Maltese texts

In determining the readability of Maltese texts, the study was not, however, able to use five of the different algorithms applied to the English texts as these tended to be language dependent. The only exceptions to the language-dependency algorithms were the Lasbarhetsindex (LIX) formula, the Automated Readability Index (ARI) and the Coleman-Liau Index (CLI). Although these last two were not tried and tested on Maltese texts, they were independent of language and syllabic characteristics (Reck & Reck, 2007; Tillman & Hagberg, 2014) and deemed applicable for a comparative analysis of the Maltese texts in question.

The three formulae taken from Tillman & Hagberg (2014) and applied as algorithms to this part of the analysis are listed here.

Equation 5.3-1

$$LIX = \frac{W}{S} + \frac{X \times 100}{W}$$

W = Total word count

S = Total number of sentences

X = Number of long words with 6 letters or more

Equation 5.3-2

$$ARI = \frac{4.71L}{W} + \frac{0.5W}{S} - 21.43$$

L = Total amount of letters, numbers, and punctuation marks

W = Total amount of words

S = Total amount of sentences

Equation 5.3-3

$$CLI = \frac{5.88L}{W} - \frac{29.6S}{W} - 15.8$$

L = Total amount of letters, numbers, and punctuation marks

W = Total amount of words

S = Total amount of sentences

All three formulas were reviewed and tested by Tillman & Hagberg (2014) and although that study was applied to Swedish and English only, Tillman & Hagberg considered them relatively reliable for processing other non-English texts. Like ARI and CLI, LIX does away with language and syllabic-dependent variables and determines readability according to character, word and sentence statistics. Other work by Anderson (1983) and Reck & Reck (2007) also confirmed the applicability of the LIX formula to different alphabetised languages other than English due to its language-independent nature.

The LIX algorithm had been applied successfully in previous work by Mifsud (2019) and was used as the main analytical tool for Maltese texts in this study. However, the ARI and CLI indices also proved practical, and their outcomes were also applied to determine if their trends reflected the same pattern as that of the LIX algorithm.

One issue that arose during the analytical process was that the study initially tried to apply an online version of the LIX, ARI and CLI algorithms to process Maltese texts, however, these online algorithms proved to be unreliable and inconsistent in their outputs. The research needed to develop a dedicated process to determine the readability of the Maltese texts accordingly. This is detailed in the process section 6.4 below (Processing Maltese texts to determine readability).

The outcomes of the LIX, ARI and CLI analysis for each of the Maltese texts used in the successive test forms were plotted on a graph for comparison using the same methodology as that applied to the English readability analysis. The CLI and ARI algorithms offered a similar outcome measure to each other that was however different to the LIX measure. A separate graphical analysis of the outcomes from the ARI and CLI algorithms was therefore plotted on the same grade-level vs year graph and then compared to the outcome trends on the LIX plot. No data synthesis or aggregation of outcomes was performed for the Maltese texts.

iii. Limitations of using readability formulas.

The application of the reading algorithms used in this study makes this part of the analysis dependent on word, sentence, and syllabic characteristics of the text to determine readability and is thus a statistical analysis of the text. It was therefore not able to shed light on variations in complexity due to the individual word items being more or less challenging for the age group.

Anderson & Davison (1986), have argued that applying readability formulas to text will deliver a statistical model of the text to which there may be a correlational relationship to difficulty level, but no direct causal inferences about individual or group comprehension may be

deduced. Similar arguments are put forward by Reck & Reck (2007) that statistical analysis of a text is not a conclusive reflection of complexity or difficulty level. However, Reck & Reck (2007) argue that LIX would be effectively reliable when applied in relative comparison to similar texts. That being said, the purpose of running a readability test on comprehension texts for Maltese and English was not intended to be the final determinant in defining test form complexities. They were applied as one component of a broader work to understand what the longitudinal trends were when considering the test form complexities.

5.3.3.2 Cognitive item demands

This second part of the analysis of extraneous cognitive load was purposed to identify variations in the mental load exerted by the individual test items and determined from a review of the questions verbs used for each test item. This collective test comparison was done using Bloom's taxonomy as the tool to categorise the cognitive demand of the different test items. Multiple cognitive taxonomies could be applied as a cognitive tool to underpin the analytical process (Dueñas et al. 2015; Newman et al. 1988), however, Bloom's proved more pragmatic in establishing a structured algorithmic system to process the data. Such systems were used by Chang & Chung (2009) and Dueñas et al. (2015) to categorise cognitive demands of question items.

More relevant to the processes used in this research, Jones et al. (2009) used Bloom's taxonomy to run a similar process, however, they used a broader scale of "high", "intermediate" and "low" to categorise the cognitive demands posed by test items. This system was applied for determining the overall cognitive demand of intermittent JLEE exams as it simplifies issues that arise when more than one interpretation is possible for the same question verb. Further details of the process are presented in the next two sub-sections.

i. Intermittent test form selection

The analysis of question verbs was conducted for each of the five subject papers during four key years 1998, 2000, 2005 and 2009 rather than for each of the fourteen consecutive sittings and a comparison was taken across these four years. In making the selection, the study determined if there had been any long-term redistribution of CL rather than a year-on-year variation.

The 1998 examinations were selected as the first set of test forms that preceded the NMC policy primarily as this is when the yearly examinations reports began to include more detailed reporting structures including projected difficulty levels and an item response analysis for each

section. 2000 and 2005 were the years that the policies were introduced, however, these were selected to offer an equidistant time gap with the last set of examinations in 2010.

However, 2009 was then decided upon as being a better choice than 2010 as it was the last year that saw all five examination papers used as part of the whole examination set— social studies was dropped from the set in 2010.

ii. Determining variations in cognitive demands

The EAU reports for each of the subjects included a projected difficulty level for each of the questions and sub-questions. This was, however, an approximation of difficulty as presupposed by the test writers. No reference is made in the reports to applied standards or mechanisms to inform or support these judgements. These projected difficulty levels were reviewed in this study's analysis section along with a further analysis of the questions' cognitive demands determined by the categorisation of each of the question verbs using Bloom's Taxonomy as a framework for analysis.

So, in reviewing the overall cognitive demands made by each test form, the study endeavoured to determine the overall picture of the CL exerted as a percentage distribution of the cognitive demand. To this end, the research drew on the framework applied by Jones et al. (2009) to categorise each of the question items according to the *"question verbs"* (2009, p. 3). The question verb informed the determination of cognitive level according to Bloom's taxonomy (L. W. Anderson & Bloom, 2001) with the main categories applied being higher, intermediate, and lower cognitive demand.

A list of question verbs and their associated cognitive levels were taken from Armstrong (2016) and Stanny (2016) to help inform the categorisation process (See Appendix B: for detailed list). These verbs were categorised according to Bloom's six main categories: Lower Cognitive Level — Remembering and Understanding; Intermediate Cognitive Level — Applying and Analysing; Higher Cognitive Level — Evaluating and Creating. The verb lists and categorisation by Armstrong (2016) and Stanny (2016) complemented each other and broadened the tool's capacity to analyse CL according to an expanded variety of verbs.

Additionally, four variations to particular question verbs were included with the appropriate category list if they were present on a question item but missing from the list. "True/False" questions, "Gist", and "Complete the sentence" (fill in the blanks) were added to the lower cognitive level. "Conjugate", referring to the conjugation of verb tenses or other grammatical conversions in language exams, was added to the intermediate cognitive level as an application of knowledge. There were no higher-order variations.

In conducting the analysis, each question and sub-question on the test form was reviewed and the question verb was identified. This was compared to the combined list of question verbs drawn from Armstrong (2016) and Stanny (2016). Those question verbs that had multiple classifications on the combined list were considered in the context of the test set and categorised under a single category. So, the term "Write", which on the combined list was listed three times under remembering, applying, and creating was, in the context of this analysis, classified solely under the remembering category of the framework. This categorisation system was applied consistently across all test forms.

Furthermore, the test forms were designed as a mix of open and closed question types and each item had a weighted score. The categorisation process therefore multiplied each question verb by the weighted score to get a truer representation of distribution according to cognitive level. This was particularly needed for open-question types that called for explanation or reasoning and carried a higher score than closed-question types that usually had a single-point score.

Once the categorisation process was complete, the weighted distribution of cognitive demands was added, and a percentage distribution determined for each of the selected test forms for each of the subjects. The outcomes were then used to establish a cognitive demand profile for that test that could be compared to inform the longitudinal analysis of change over the fourteen years in question. These profiles were not definitive in their implications but were useful to establish an approximation of general trends over the long-term.

iii. Limitations of analysing cognitive item demands

Two particular areas limited the detailed determination of cognitive load. Firstly, in reviewing the questions to determine their type according to Bloom's taxonomy, an initial attempt was made to use all six categories, however, ambiguities arose in determining the exact category of the cognitive demand posed by the question. For the most part, these ambiguities occurred when determining if the statement assessed remembering and/or understanding. The intent of the examiner was not as distinguishable as the categorisation of the action verbs using Bloom's taxonomy suggests. Using a more general categorisation — higher, intermediate, and lower CLs — simplified the process of classification and reduced the ambiguities. As most of the uncertainty lay in deciding whether the demand was for remembering or understanding, grouping them as a lower cognitive level facilitated that process.

Secondly, the analysis was structured to maintain objectivity and sustain as straightforward a categorisation process as possible. To this effect, interpretation of examiner intention was

avoided whenever possible, and the items were categorised based solely on the question verbs used. This however misrepresented some of the questions which (especially in mathematics) had a deeper evaluative intention based on multi-step processes. The concept of "Work out" in particular would be used for single-step processes but could also represent a multi-step process on the examinations. By consistently categorising it as a single item for all test forms, the analysis reduced all multi-step processes to a single category.

5.3.3.3 General format and structures of the items

The final analytical strand investigating extraneous cognitive load looks at the general format and structures of the test forms and associated items.

Gillmor et al. (2015), have argued that the management of visual aids, signalling, weeding and sequencing (2015, p. 6) plays an important part in reducing extraneous CL on test forms. They identify work by Hitch (1978) and Kettler et al. (2009) recognising the impact of textual and numerical paring, and with Miller's (2011) consideration of the importance of aesthetic presentation. Gillmor et al. (2015) identify seven key strategies that can be controlled to modify the extraneous CL on an assessment: Translation; Visual Aids; Signalling; Weeding; Sequencing; Aesthetics; and Numerical simplicity (2015, p. 6). Moreover, their work suggests signalling, weeding and aesthetics may have had a greater effect on making test forms more accessible to test subjects (2015, p. 15).

In identifying longitudinal variation in difficulty levels for the JLEE, a comparison of test forms from pre- and post-policy introduction periods would give an idea of any changing techniques that may have been used to influence accessibility. Comparing the set of JLEEs, the analysis compared the format and characteristics for the initial and final years — 1997 and 2010 (2009 for social studies). The analysis, however, also included 2004 as a midway transitional point, due to it falling after the introduction of the NMC (2000) and before the introduction of the NCF (2005) while also coinciding with the midpoint between 1997 and 2010.

The work by Gillmor et al. (2015) allows a loose framework to be established that can be used to underpin a comparative review of the test forms for each subject. The framework was structured around the table proposed by Gillmor et al. (2015, p. 6) but looked to identify changes in strategy.

Table 5-1 Analysis framework for the general format and structures of test forms.

STRATEGY	ANALYSIS WILL LOOK FOR A CHANGE IN…
TRANSLATION	…word count and language level.
VISUAL AID	…the number of diagrams to support spatial information.
SIGNALLING	…the type and number of signals and cues to focus attention
WEEDING	…the number of visuals and text extraneous to the item construct
SEQUENCING	…whether or not scaffolding and logical ordering of questions and sub-questions are applied
AESTHETICS	…the organization of graphics and overall test flow (clutter of text and use of white space).
NUMERICAL SIMPLICITY	…the use of smaller, simpler numbers when values are construct-irrelevant.

Source: Gillmor et al. 2015, p. 6

The work by Gillmor et al. (2015) was established on assessment in mathematics and was appropriately applied to the mathematics test forms used for this study. For the other four subjects, however, the relevance of each strategy varied with strategies of sequencing and numerical simplicity proving irrelevant as the constructs for these four exams were designed to test distinct knowledge items from different subject areas. Furthermore, strategies associated with visual aids were considered more relevant to mathematics and social studies (geography and history), but not as relevant to religion, English or Maltese test forms. Nevertheless, consideration of the use of visual aids was conducted with the latter three sets of test forms. Table 5-2 below lists the test forms to which the strategy analysis was applied.

Table 5-2 Applicability of strategies to different test forms

Strategy	Applicable to
Translation	all 5 subjects
Visual Aid	All 5 subjects
Signalling	all 5 subjects
Weeding	all 5 subjects
Sequencing	Mathematics
Aesthetics	all 5 subjects
Numerical simplicity	Mathematics

Similar to the previous section, this framework allowed general insights into changes that would have affected the test forms based on a common test construct but will be tentative in its interpretations rather than definitive in its conclusions.

5.3.3.4 Overview of the mechanisms employed for cognitive analysis

The main purpose of this stage of the analysis was to identify, in a cursory manner, any extraneous additions or reductions that may have taken place to the test forms to change the cognitive demand exerted on the test subjects making the papers more or less difficult for successive cohorts.

Readability variations, question verb analysis, and the structure and format reviews were intended to act collectively to inform an overall trend analysis of difficulty level variations of the test forms based on an approximate determination of variations in the extraneous CL. The statistical analysis of difficulty levels that followed would then give a clearer, more definitive measure of any such changes. Both sets of analysis would then work in a complementary manner to inform the interpretation of the overall analysis of student outcomes on the Junior Lyceum Examinations.

5.3.4 Statistical analysis — Trends in psychometric characteristics

The second part of the analysis of context is structured to determine latent trend variation in test form complexity. It presents an analysis designed to determine longitudinal trends in quality and standards using a graphical methodology to examine relative variations of psychometric characteristics. The cognitive analysis described in the previous sections was intended to be used together with the statistical analysis of the Discrimination (D_i) and Facility (F) indices described in this section.

As postulated in the literature review, D_i and F indices can be used to reflect on the quality of each test item even though they might not adequately show the degree of cognitive challenge on an assessment. If each examination is therefore taken to be the set of all test items, it should then be possible to determine information about the quality of each exam through a collective analysis of the associated D_i and F for all items on the test. Such an analysis was facilitated by the EAU reports presenting both F and D_i for each test item from 1999 to 2010 (Curriculum Department & Educational Assessment Unit, 1999 - 2010).

This part of the analysis is organised into two subsections. The first discusses a statistical analysis that considers variations in the mean for D_i and F over the 12 years, while the second looks at a graphical analysis of trend variations exhibited on D_i vs F plots for the same data. The statistical analysis alone would give an idea of changing quality, but the graphical analysis renders a more detailed understanding of common traits and shifting patterns for these psychometric measures and reflect underlying shifts in examination standards.

5.3.4.1 Variation in the mean of D_i and F

Statistical analysis for each set of psychometric data was conducted to look at trends for both D_i and F. This required plotting the annual statistical mean for each of these variables as a percentage (%) against time (t) thus quantifying any longitudinal changes to inform an initial understanding in this regard. Five graphs were prepared, one for each subject with the respective plots of D_i and F presented on the same axis. A trendline was added to each of the plots with the gradients of each trendline reflecting on changing quality for those subject sets.

5.3.4.2 Psychometric characteristics for parallel test forms

As the statistical analysis was established on the collective mean of the psychometric measurements it could not reflect any changes in the variance of the item characteristics. This statistical analysis was therefore followed by a graphical analysis of D_i vs F plots organised in a comparative array and intended to further elaborate on any longitudinal variations in the quality and standards of the examinations.

The concept of D_i vs F plots and the associated curve for an ideal test has been discussed in the literature sections and is being proposed as a comparative tool that would allow clearer insights into longitudinal variations in quality, difficulty levels and standards. It is important to re-emphasise that quality reflected in the psychometric characteristics of test items is context-dependent and varies according to the design purpose and the scoring interpretation of the exam (American Educational Research Association et al. 2014, p. 38; Ebel & Frisbie, 1991).

The EAU stipulated a need to have balanced exams that were neither too easy nor too difficult in order to *"discriminate between the different ability levels it was intended to measure"* (Curriculum Department & Educational Assessment Unit, 2005, p. 15). This underscored the EAU's intention to maintain a standard for test construction that offered controlled sets of test items. It also indicated an intention to have some form of continuity and consistency in examination quality and standards for these exams. Considering that the subsequent test forms were drawn from the same constructs and administered to similar cohorts then, as long as the quality and standards for exam setting remained the same, the outcomes should give similar sets of psychometric measurements from one year to the next.

Comparative arrays of D_i vs F plots

The tables presented in the EAU reports listed D_i and F in two columns for each question item and although it allowed the possibility of identifying how well individual question items conformed to the criteria, it did not give a clear understanding of the quality of the tests as a

whole. To render a clearer understanding of year-on-year trends in the psychometric data, the study used each set of test item data to produce the graphical plot of Di vs F.

The aggregation of these item analysis data into an array of D_i vs F plots offered a collective overview of the quality of each examination set and a means of determining longitudinal variations. Although this can be done by simply analysing the change in the statistical mean associated with the psychometric data the array of plots gave a better idea of quality variation on the examinations through the shifting distribution of points across the plot area and approximation of those trendlines to the ideal-test curve.

Distribution characteristics of D_i vs F plots

There is a second consideration to be made here regarding the actual plot distributions and associated trendlines. The stated EAU objective of having a fair balance of controlled questions could be used to support an argument that the plots of D_i vs F should render an ideal-test curve as discussed earlier (Karelia et al. 2013; Sim & Rasiah, 2006). However, even though this latter statement may not signify that the distribution of the plot would necessarily approximate an ideal-test curve, the main argument remains that the EAU intended to retain a similar level of quality and standards from one year to the next. This implies that the series of annual plots of D_i vs F for parallel test forms should exhibit similar distribution characteristics and trendlines if they have the same standards and quality characteristics, irrespective of what the general shape of those trendlines might be.

5.3.4.3 A comparative framework for analysis

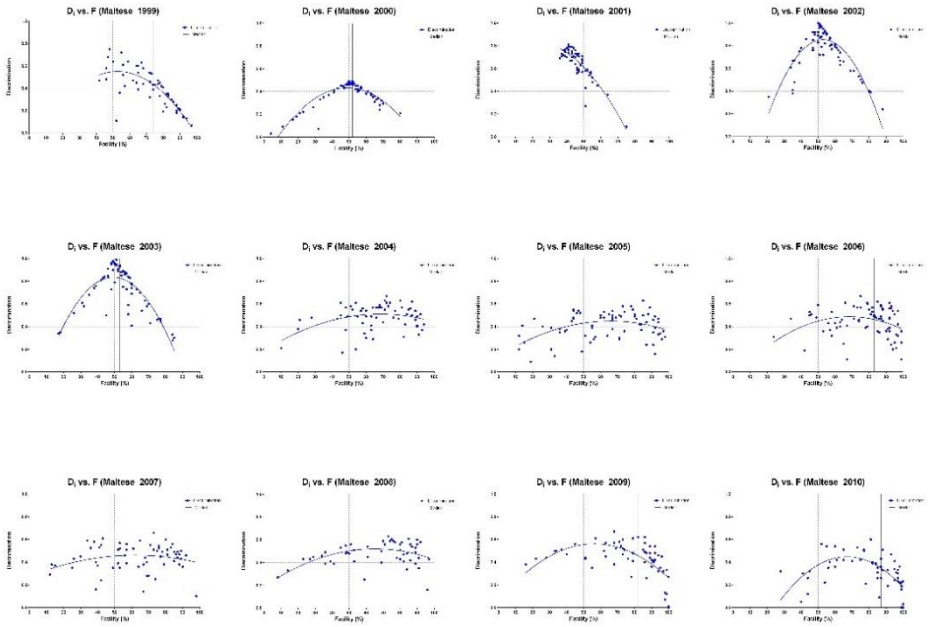
To get a better understanding, of the longitudinal variations in each of the test sets, this study needed to develop a tool to exploit the expected continuity in D_i vs F plots.

The argument posited in the previous section proposed that test forms that maintained the same design standards and were based on the same test constructs would have a similar level of quality and offer similar plots of D_i vs F for similar cohorts. This would conversely imply, albeit not categorically, that for tests derived from the same construct and administered to similar cohorts, then any plots that displayed similar distribution characteristics should reflect relatively similar quality. This study subsequently made use of the graphical plots of D_i vs F for each of the examination sets from 1999 – 2010, comparing them longitudinally for subject-based commonalities or variations, and cross-sectionally to understand trends across the subjects.

The psychometric data used to plot D_i vs F was taken from the EAU reports (Curriculum Department & Educational Assessment Unit, 1999 - 2010). The statistical item analysis for 1997 and 1998 could not be included as the process was begun by the EAU in 1999. Furthermore, this data was only available as a hard copy and needed to first be digitised before being processed and plotted (Chapter 6 Digitisation Process). These plots were then organised into an array of 5 subject rows and 12 yearly columns. This allowed longitudinal comparison of the trendlines across each row and facilitated three possible comparative interpretations for each subject set over time — the degree of correspondence or continuity; any abrupt changes; any gradual variations.

To be able to present the array in a manner that could be read and compared appropriately, each subject row was assembled onto a single page as shown below in Figure 5-2 allowing the plots to be presented by subject as a linear sequence and compared.

Figure 5-2: Example of a subject row for longitudinal comparison (Maltese 1999 - 2010)



5.3.4.4 Analysis of a graphical array

This analysis was underpinned by visual qualitative observation of the general patterns of the plots against time to identify any common variations over shorter or longer periods considering plot densities and the associated trendlines. In making this comparative analysis, the study first considered the distribution density of the plotted points with variations in those densities reflecting on both the difficulty level and discrimination power and subsequently on the quality of those exams. The second part of the comparative analysis then looked at the shape of the trendlines and the approximation of those trendlines to the ideal-test curve (D_i =

0.4, F = 50%), selected to match the EAU criteria. The ideal-test curve established a comparative background against which to describe the characteristics of the individual plots and identify commonalities or differences.

The analysis was conducted through three comparative processes. In the first instance, the analysis considered subject-specific longitudinal changes. The second was the cross-sectional analysis identifying yearly differences and commonalities between the different subjects. The third instance considered a comparison of the five annual exams as a group — an aggregated longitudinal comparison of the plots — and was able to identify unusual distinctions for different years. This latter aggregated comparison came about following preparation for the first two instances and reflected possible institutional factors that impacted all the examinations similarly.

As part of the analysis, the cross-sectional and aggregated longitudinal comparison are merged as a lot of the observations seemed to overlap and it made more contextual sense to discuss them at the same time.

5.3.4.5 Limitations of statistical analysis of psychometric characteristics

Although the statistical variations in the psychometric measurements were quantifiable, the earlier discussions regarding D_i and F posited that these values were not absolute in their nature and could not therefore offer an exact measure of change. A similar argument is recognised for the comparison of the arrays and the accuracy of their interpretation.

It should also be restated that the study recognised that in making such longitudinal comparisons, the multitude of influences that have an impact on outcomes cannot be considered in their entirety (Coe, 2010; Crisp & Novaković, 2009; Dorans et al. 2010; Newton, 1997; Patrick, 1996). The principal purpose of this part of the analytical process remained focused on determining and articulating observations associated with shifting trends in difficulty levels and discrimination power of the test forms and subsequently shifting trends in the quality and standards.

This part of the analytical process therefore worked to render a combined picture of trends from both the statistical and comparative analysis to understand the general variations in trends. To this end, the analysis of change in the respective statistics functioned as the main instrument for the interpretation of the psychometric data and was enhanced by a comparative analysis of the respective graphical arrays for each subject.

5.3.5 Summary

This section was introduced to understand part of the contextual dimension related to the quality framework (i-p-o-c) as described earlier. It is linked to the output dimension by rendering a contextual background against which to analyse the record of results described in the next section. An analysis of the examination reports was structured to determine the level of construct continuity and consistency together with test form difficulty levels and discrimination power.

For the analysis of the outcomes of the Junior Lyceum examinations to maintain comparative validity, the test constructs needed to retain continuity and consistency over the period in question. The analysis of context was therefore structured to investigate possible effects of some of the intrinsic and extraneous factors that could have affected student outcomes, analysing possible variations in difficulty levels of the test forms or changes in the test constructs. Variations in these characteristics would have had a direct influence on outcomes and a trend analysis was used to identify changes to these contexts over both the short and long-term. Data resources associated with understanding any associated germane CL were not available and did not form part of the analytical framework.

5.4 Data analysis – Identifying fluctuations in examination results

In looking to understand the impact of the NMC and FACTS policies on student outcomes the results of a national annual examination administered to the different yearly cohorts were processed and analysed. The examinations were criterion-referenced and viewed informally as *"an important benchmark in our educational system"* (Curriculum Department & Educational Assessment Unit, 2005, p. v). This part of the methodology presents a structured investigation of the quality framework's output dimension and was directly interpreted against the results of the contextual analysis described in the previous section.

It is understood that results from assessment and evaluation have become critical tools in establishing how well school systems are performing and providing feedback on the outcomes and attainment aspects of quality education (Santiago, n.d., p. 8). An OECD report reviewing common policy challenges for improving school outcomes stipulates that system performance monitoring can be established on a variety of national assessment programmes. They further recognised that there is *"Greater reliance on evaluation results for evidence-based decision making."* (OECD, 2013, p. 13). These monitoring processes that work to inform decision making can effectively draw evidence from student achievement data (Iwu et al. 2018; White, 2020).

A time series analysis of the results record was therefore used to investigate fluctuations in student achievement over the years. The time-based comparative characteristics together with the establishment of fixed points defined by the policy implementation allowed insights into ongoing developments and subsequent impacts on outcomes. As explained earlier, consideration was given to factors that would have affected those outcomes through variations to the intrinsic or extraneous CLs.

5.4.1 Defining the periods for analysis

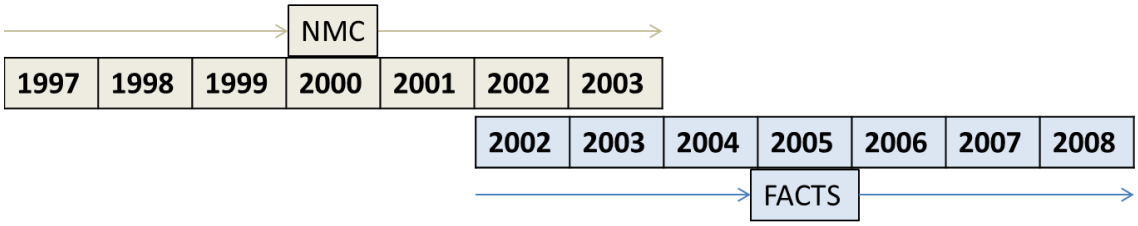
An analysis of effect due to policy change should span a singular point in time—the year at which the policy was introduced. The analysis of outcomes focused on a range of ±3 years on either side of the policy introduction to identify any fluctuations or patterns, within the context of that period. Any changes detected within the three years following the introduction of the policy, when compared to the three years preceding, can then be used to argue a possible policy role in affecting the change even if, as stated by Collins (2006), the argument will not be final in its conclusions.

Coe et al. (2008) stated that when comparing educational standards across time, the longer the comparison period the less meaningful the interpretation. Patrick (1996) and Newton (1997) argue that this is due to changing contexts associated with changing culture and technology. However, it is also accepted that over shorter periods, there will be stronger implications that can be drawn from the analysis (Crisp & Novaković, 2009; Newton, 1997, p. 227; Patrick, 1996). This implies that the validity of comparative analysis has an inverse proportional relationship to time (in years), and the exact relationship is furthermore contextually dependent on other social processes. Subsequently, comparative analysis on a shorter timescale tends to be more relevant than longer-term considerations.

In considering longer-term effects, Gray, Goldstein, & Thomas (2003) have argued that long-term data would allow an analysis of policy effects over a longer period and be able to consider belated effects of in-school policies (2003, p. 87). Consideration needs to be given to the possibility of time-lagged effects that may begin late within this timeframe and continue beyond.

The policies in question were initiated in 2000 and 2005 and the data collected ranged from 1997 to 2010.

Figure 5-3 Period leading to and following the NMC and FACTS policy introductions



The periods outlined in Figure 5-3 are being taken to be more relevant than other years that are further away from the policy introductions, but with that in mind and the availability of additional data from other years pre-1997 and post-2008, a complete time series chart was still constructed and scrutinised for other anomalies that may have taken place during other periods. This extended time series also allowed a visual comparison of periods as clusters that could then be compared to each other and reflect on the overall trends over protracted periods of time.

It should be noted that the authors of the examination reports had also included the pass-fail rates for previous years dating back to 1981. These values from the reports were used in the study to extend the plot for pass-fail rates backwards to 1988 to get a broader picture of fluctuating trends over the previous decade as well. 1988 was chosen as a cut-off date as that was the year that the ministry had introduced social studies and religion as part of the examination set (Grima et al. 2008, p. 29) and made the study more comparatively relevant by including all five subjects across the board.

5.4.2 The use of achievement results

In working to respond to the third RQ and understand the impact of the two policies on outcomes and quality of education, an analysis of the record of results was conducted to identify any longitudinal variations in achievement. The introduction dates of the NMC and FACTS policies were then overlayed on the outcomes of this time series to establish points of reference to support before and after comparisons.

Moreover, this time series was reviewed together with the contextual analysis to try to determine if any signs of impact were a result of the introduced policies or other affecting factors. The analysis of context discussed in the previous sections considered variations in test constructs and test form complexity to elaborate on possible affecting factors that might have had an impact on student achievement.

In order to establish identifiable variations in the record of student outcomes, a multilevel analysis was conducted of the yearly achievement results that considered both overall and

subject-specific outcomes. The study used four different processes: overall pass-fail rates; aggregated grade averages; varying grade proportions using banded grade percentages; and an analysis of single/combined subject failures. Furthermore, due to fluctuating cohort numbers over the years, this analysis relied on percentage measures of the associated variables.

Pass-fail rates and aggregated grade averages

The first two methods were used to understand variations in overall student outcomes first using the EAU's success criteria and then an aggregation methodology that determined the overall average score for each student.

Although the pass-fail rates reflected the year-on-year variations in attainment for Year 6 students, these results alone were considered to offer a skewed picture of changing standards due to the conjunctive success criteria (Douglas & Mislevy, 2010; McBee et al. 2014) applied by the EAU. The research required a further, more comprehensive understanding of the data that would offer a more holistic analysis of the outcomes by combining student results in a compensatory model (Douglas & Mislevy, 2010; McBee et al. 2014). This system was intended to reduce any skewing of outcomes inherent in the application of the EAU's conjunctive criteria.

Kickert et al. (2021) applied a similar grade point average *"GPA as the weighted average of the final grades"* (2021, p. 1045) to compare conjunctive and compensatory performance outcomes for the same cohorts. They have argued that the compensatory model results in greater student achievement compared to the conjunctive model, due to the conjunctive model requiring the completion of all hurdles without a single failing score.

Analytical methods using aggregated grade averages were therefore considered to offer a nuanced picture of variation in overall achievement than the initial pass-fail rate analysis.

Grade proportions, banded grade %, and single/combined subject failures

The final two methods focused on determining if there was any weighted influence of the different subjects on the overall achievement of results. These were subject-specific analyses and were introduced at a later stage when it became apparent that English and mathematics had a disproportionate influence on student success rates. Banded grade percentages and single and combined subject failures were developed to investigate the degree of influence that each subject had on the overall achievement results.

5.4.2.1 Pass-Fail rates

The pass-fail rate analysis was a time series analysis of percentage pass-fail ratios determined using the EAUs criteria for successfully passing the JLEE. These conjunctive criteria required that a student achieve a grade of C or higher in all five exams (unless exempt) to be given a pass. The data for this analysis was drawn from the record of results compiled through the digitisation process and spans all 14 years under review.

An additional procedure was included in the graphical analysis that made use of the difference between the percentage of passes and fails (Δ% = %pass - %fail). This difference is directly related to the pass-fail ratio and does not add any more analytical details to those rates of change. However, it does add emphasis to the year-on-year changes in achievement that were taking place, visually magnifying trend variations on the graphical presentation for easier observation. Furthermore, the rate of change of Δ% can be taken as an indicator of improvement when comparing gradients over similar periods, with positive, negative, or constant variations of slope indicating changes in improvement rates.

The analysis used the Δ% rates to compare the decade before and after the introduction of the NMC. Data for the success percentages were presented in the EAU reports going back to 1988 (Curriculum Department & Educational Assessment Unit, 2010, p. 10). This data was used to extend the analysis to the nine years before 1997 and thus extend the analysis of the pass-fail rates. Segmented regression analysis (Lagarde, 2012; Wagner et al. 2002) was then used to compare the annual rate of change of the difference Δ%, for the decade before and after the introduction of the NMC in 2000. Specifically, the analysis considered the rate of change of Δ% during these two periods to understand if there was any difference to that effect. Variation to the rate of change of Δ% would reflect on changes in outcomes and achievement and could be argued to be a further reflection of impact on outcomes.

Lagarde (2012) has presented arguments supporting such an analysis using *"retrospective longitudinal data"* (2012, p. 77) when baseline surveys were not an option. She further argues the application of regression analysis of the pre- and post-intervention data to support arguments related to impact. Wagner et al. (2002) placed similar importance on segmented regression analysis to determine the impact of interventions when controlled experimentation was not an option. Work by Marston (1988) used a method of comparing gradients similar to what is being proposed in this section, however, the context was a smaller-scale reading intervention. Nevertheless, Marston applied a comparative before and after regression

analysis establishing the pre-intervention as the control against which the post-intervention was to be compared (1988, p. 16).

The regression discontinuity analysis used in this study required that the rate of change be determined from the gradient of the Δ% trendline for the pre- and post-NMC periods with any changes indicating a variation in improvement rates. This was done for the decade before and after the introduction of the NMC in 2000 and these rates of change were compared.

5.4.2.2 Aggregated grade average

As stated earlier, the criteria established by the Department of Education failed students if they received D, E, or abs on any one of their 5 assessments. This conjunctive rigidity in determining the outcome for each student tended to skew overall achievement (Grima et al. 2008, p. 69) by amplifying the effect of a single non-passing result (McBee et al. 2014). A student receiving four A's and a D would inevitably fail despite having an aggregated grade average of B.

In working to gather more insight into the impact of the policies on achievement, consideration was given to try and reduce the effect of this inherent skewing on overall achievement. The overall average grade was calculated for each student and the pass or fail was determined by results above or below an average outcome of C. To do so, each grade was converted to a numeric value A-1, B-2, C-3, D-4, E-5, abs-5, x (exempt)-no value and the mean was then calculated assigning a pass or fail to each student according to that average value.

In determining the averages, the calculated mean result was rounded to the nearest integer. This meant that average scores less than 3.5 were rounded down to 3 (grade C) while those equal to or above 3.5 were rounded up to 4 (grade D). Similar criteria were applied to all other grade bands.

It was initially thought that this process might bias any comparison with the conjunctive EAU criteria pass-fail rates from the previous section by shifting the pass rate beyond the C grade (represented by a mean score of 3) to a point between grade C and grade D, as would be represented by a mean score of 3.2 or 3.4. However, this averaging process actually established appropriate banding ranges for each of the grades that were not possible through the application of a compensatory model, vis-à-vis the applied EAU criteria, and allowed a different perspective on actual student achievement. Table 5-3 shows the banding established for the compensatory model discussed here.

Table 5-3 Banded grade score averages.

A	B	C	D	E
$1.0 \leq \bar{x} < 1.5$	$1.5 \leq \bar{x} < 2.5$	$2.5 \leq \bar{x} < 3.5$	$3.5 \leq \bar{x} < 4.5$	$4.5 \leq \bar{x} \leq 5.0$

In working to reduce the skewed nature of the outcomes this process was able to determine if there was an overall change in the aggregated grade averages over the fourteen years being considered. Similar to the pass-fail rate analysis described in the previous section, once the aggregated grade averages were calculated, graphical tools were used to analyse those scores and an analysis of the difference and the rate of change of that difference was used to determine longitudinal variation.

Norm-referenced determination of grade boundaries

It needs to be noted at this stage that up until 2002 the EAU used norm-referencing to establish grade boundaries for grades A and B so that the top 5% (approx.) got A's, the next 20% (approx.) received Bs, and any raw score over 50% but outside the cohort's top 25% was assigned a C grade with the rest being given Ds or Es (Curriculum Department & Educational Assessment Unit, 1997 - 2002). The policy for grade assignment after 2002 is unclear and was no longer reported in the EAU reports. This system for establishing grade boundaries would inevitably impact the aggregated grade average such that the true average could not be determined accurately without using the raw scores.

However, applying such an averaging process could still work to reduce the bias introduced by the EAU's minimum requirements as the cut-off mark for achieving a C grade remained a 50% raw score for each subject, meaning that the number of those successfully passing each subject should have remained similar. The validity of any conclusions drawn from such an analytical process should certainly be reduced but could still be considered to shed light on overall student achievement.

5.4.2.3 Grade proportion distributions - Banding grade scores

The third part of the proposed analysis of results looked at the proportional distribution of grades across the fourteen years for each subject and overall. Shifts in achievement levels would be reflected as variations in the proportion of passing grades (A's + Bs + Cs) compared to failing grades (Ds + Es) for each subject.

The norm-referenced system used by the EAU to establish grade boundaries meant that comparing grade distribution longitudinally would not render any meaningful analysis and invalidate comparisons with the subsequent years (after 2002) that did not use the same

system consistently. Furthermore, the lack of clarity of what new policies were adopted started in the 2003 EAU report making any analysis of the proportional distribution of grades unreliable (Scott, 1990).

However, and in order to supplement the analysis of achievement, it was possible to band the graded scores into 2 groups considering the banding of A's + Bs + Cs as a single group and Ds + Es as the second group. This was possible for three main reasons: first, those achieving a raw score of 50% or more received an A, B or C grade but not otherwise; secondly this criterion was applied consistently across the 14 years; and third, there was no overall norm referencing applied to the raw or final grade scores. The distribution of banded scores by subject was then applied to reflect on overall achievement specific to each subject and consider how that achievement varied over time.

Similarly, this banding technique was applied to an aggregate count of the overall grade results for all subjects counting the banded scores. This analysis intended to introduce a further comparative dimension of the proportional distribution of passing grades to failing grades for each subject and overall.

5.4.2.4 Single and combined subject failures

In the process of analysing the data, it became clear that English and mathematics had a disproportionately stronger influence on the success rates of students than the other three subjects. To get a clearer picture of the degree of influence that each subject had on the overall student outcomes, part of the analysis considered the impact of single-subject failures on the pass-fail rates. An Excel algorithm[4] was run on the array that had been created by converting grades to numerical values (as discussed in 5.4.2.2 above) and, for each of the examination sets, the number of cases that failed that subject alone was counted. As the EAU criteria stipulated that any single grade less than "C" constituted an overall failure, this avenue of analysis allowed the study to understand how many single-subject failures resulted in overall failures and if such situations were cross-sectionally consistent or not. Cross-sectional consistency would indicate that no one subject had a greater impact on pass-fail rates than any other subject.

---

[4] *Example of Excel algorithm for English (Column U):*
=COUNTIFS(S:S,"<=3",T:T,"<=3",U:U,">3",V:V,"<=3",W:W,"<=3")

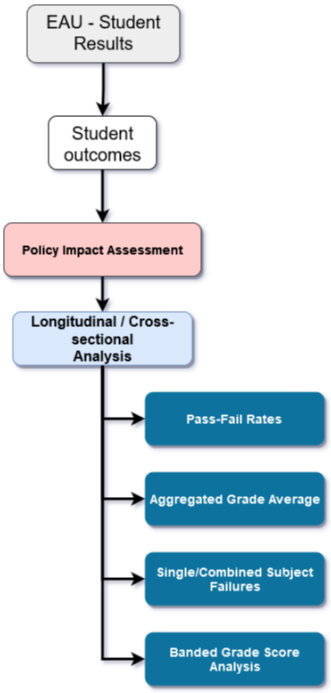
Similarly, and using a similar algorithm, a combination of subject failures (SS & ML; SS & RL; ML & RL; SS & ML & RL; and EN & MT) that resulted in students being unsuccessful in their JLEE were also processed.

5.5 Summary

This section of the analysis presents a framework for investigating that responds to the third RQ and established the relevant periods to be considered together with a set of four time series tools (Figure 5-4) to determine policy impact on outcomes.

It relates to an understanding of quality by looking at the output dimension of the quality framework (i-p-o-c) and links to the analysis of context described in the previous section. However, this section of the analysis also adds further information to the contextual dimension by looking at subject-specific influences on outcomes. Cross-sectional comparisons between these subjects were purposed to understand consistency in progress across the years for all subjects and how they influenced achievement.

Figure 5-4 Outcomes analysis flow diagram



The data set for this part of the analysis was drawn mainly from the record of results kept on file at the EAU and, as stated earlier, constituted *"secondary data analysis"* (Hakim, 1982; Logan, 2020).

The concluding chapter of this methods section reviews the acquisition and conversion processes undertaken to digitise and validate the different sets of recorded data.

6 Digitisation Process

6.1 Chapter overview

As part of the framework of mechanisms, tools and procedures developed for this study, this chapter is structured to explain the initial groundwork required to collate, digitise, and prepare the JLEE data for processing and analysis. In so doing, this work addresses part of the first research sub-question aimed at aggregating and preparing the data and reporting records for analysis. More specifically, the discussion that follows explains the step-by-step processes used to digitise the student record of achievement on the JLEE and parts of the EAU reports.

The first part of this chapter reviews the data sources being used for this study, their accessibility, and the format in which they were made available.

As the analysis of the EAU reports and the records of achievement was based on a graphical and statistical analysis of data available only as hard copy printouts, the second section explains the step-by-step digitisation process, quality control mechanisms, and measures taken to ascertain data integrity of the digitised record.

The third part describes the structuring of a readability system applied to the Maltese texts. Following a scanning and OCR process, the English comprehension text then required a straightforward analysis of readability, however, the Maltese comprehension texts did not have readily available readability tools.

The final two sections deal with processes for determining a system for comparing cognitive item demands on different papers and the digitisation process used to render the item analysis machine readable.

6.2 Data sources and digitisation

This research centres on an analysis of three official sets of government documentation: the policies (MEYE, 1999, 2004a), the record of JLEE results (Educational Assessment Unit, 1997–2010) and, the formal reports associated with each examination set (Curriculum Department & Educational Assessment Unit, 1997 - 2010). The three policies are available through the Ministry of Education's 'Policy Documentation Archive' (n.d.) while the records of results and the associated reports were gathered directly from the EAU offices.

6.2.1 Records of Examination results

For the period between 1997 and 2010, there was a single examination cycle every year and the results were stored in two bound documents (Educational Assessment Unit, 1997–2010) and sorted according to a fixed set of criteria. The source records were organised in the two bound volumes so that the public schools in Malta were all listed first, followed by the public schools in Gozo and finally all the non-public schools. For each examination cycle, the two bound volumes were categorised A-S and S-Z and together the two volumes recorded the results for the same set of schools across all the years.

6.2.2 Digitisation of records – a necessary step

As the JLEE outcomes and results data (Educational Assessment Unit, 1997–2010) were not readily available in digital form, it became necessary to first digitise these records before proceeding with any form of machine-based analysis. This was done through a digitisation process that required scanning, and optical character recognition (OCR) processes followed by quality and integrity checks. Similarly, the examination reports (Curriculum Department & Educational Assessment Unit, 1997 - 2010) were only available in hard copy and required digitisation of the different relevant sections.

The processes described in the following sections were applied in the digitisation of the information and data record. Although some of the procedures are standard steps in most digitisation processes (scanning/imaging and OCR), other steps required a customised solution depending on the nature, format and structure of the data being processed at the time.

Similarly, quality and integrity procedures required customised solutions due to the particular format and structure of the digitised data being scrutinised. Any incorrect conversion or slippage would skew the data array and consequently distort any analysis, so accuracy was supported through rigorous error-checking mechanisms within the process that confirmed each data record was correct and aligned.

6.3 Digitising the record of results

The preparation process required the conversion of printed records to a digital format to be able to scrutinise the data using electronic spreadsheets and graphing tools for statistical analysis. However, the conversion was not an end in itself and required integrity, quality, reliability, and validation checks to ascertain that each converted item of data was accurate and maintained its position in the digitised array. The initial data preparation was therefore structured around maintaining accuracy in the digitisation process to produce a true digital record of the examination results data and the associated item analysis from the yearly examination report.

6.3.1 Format of the recorded results.

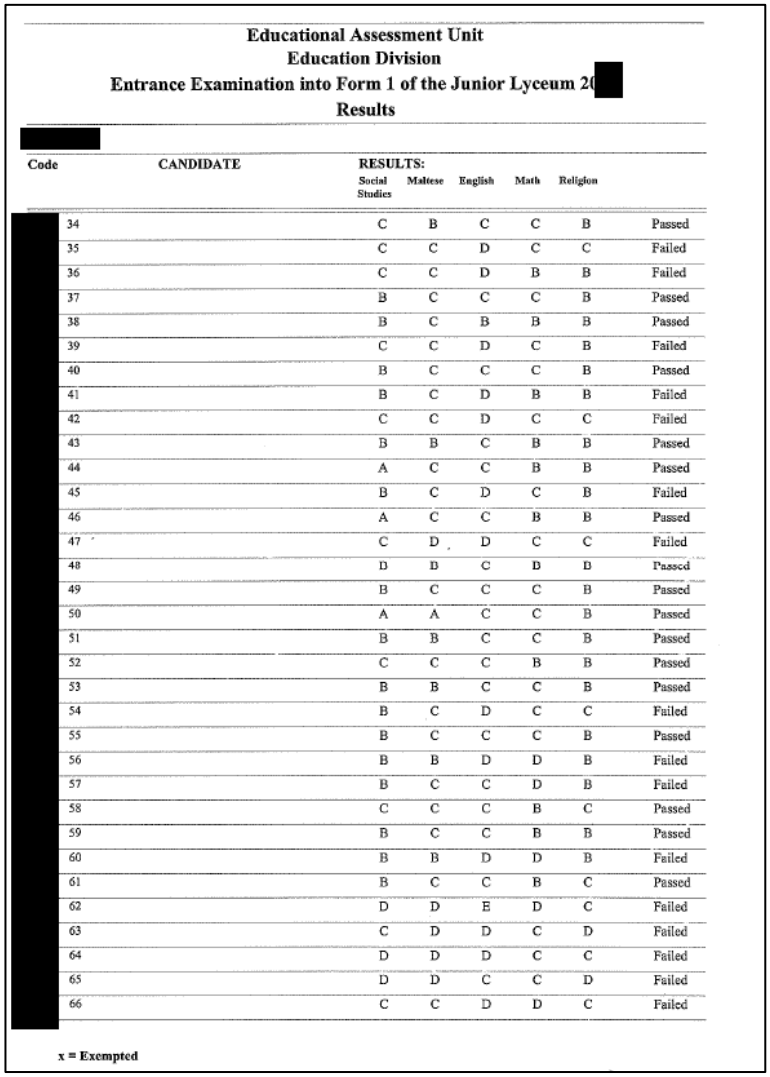
Once the MEDE granted permission, and in coordination with the personnel at the ministry's assessment unit, each page of the two-volume record of results was scanned to PDF. There were two bound volumes for each year, which resulted in two separate PDF files for each examination cycle, and as the examinations were administered once a year, there were 28 PDF files spanning the fourteen years (1997 – 2010) that needed to be digitised and processed. Each page consisted of an array as shown in Figure 6-1.

This array consisted of eight columns with the first two representing the school and student ID (combined), and the candidate's name, respectively. The subsequent five columns were the grade results for each of the five subjects (social studies, Maltese, English, mathematics, and religion) and the last column displayed the overall outcome (Passed – Failed) for each candidate. For the last examination cycle (2010) the administrators dropped the social studies examination so that the last examination cycle was based on the remaining four subjects.

The results were formally recorded in the source records on an A - E grade scale across five columns (one column for each subject) with "Abs" representing absentees and 'x' representing exemptions. A final column had the recorded outcomes (Passed-Failed) for each row.

A single row of data comprised of the results of each of the subjects, the resulting outcome, and their association with the individual candidate, and shall be referred to as an "individual data set". Whereas the complete set of data comprised of all individual data sets for that examination cycle shall be referred to as "the data set" or "the complete data set".

Figure 6-1: Scanned document showing discrete grade results and outcomes for individual students



Educational Assessment Unit
Education Division
Entrance Examination into Form 1 of the Junior Lyceum 20[redacted]
Results

[redacted]

Code	CANDIDATE	RESULTS: Social Studies	Maltese	English	Math	Religion	
34		C	B	C	C	B	Passed
35		C	C	D	C	C	Failed
36		C	C	D	B	B	Failed
37		B	C	C	C	B	Passed
38		B	C	B	B	B	Passed
39		C	C	D	C	B	Failed
40		B	C	C	C	B	Passed
41		B	C	D	B	B	Failed
42		C	C	D	C	C	Failed
43		B	B	C	B	B	Passed
44		A	C	C	B	B	Passed
45		B	C	D	C	B	Failed
46		A	C	C	B	B	Passed
47		C	D	D	C	C	Failed
48		D	D	C	D	D	Passed
49		B	C	C	C	B	Passed
50		A	A	C	C	B	Passed
51		B	B	C	C	B	Passed
52		C	C	C	B	B	Passed
53		B	B	C	C	B	Passed
54		B	C	D	C	C	Failed
55		B	C	C	C	B	Passed
56		B	B	D	D	B	Failed
57		B	C	C	D	B	Failed
58		C	C	C	B	C	Passed
59		B	C	C	B	B	Passed
60		B	B	D	D	B	Failed
61		B	C	C	B	C	Passed
62		D	D	E	D	C	Failed
63		C	D	D	C	D	Failed
64		D	D	D	C	C	Failed
65		D	D	C	C	D	Failed
66		C	C	D	D	C	Failed

x = Exempted

6.3.2 Preparation for analysis - Digitising the records of examination results

The process of digitisation required that the documents be first scanned into a set of Portable Document Format (PDF) files. Scanning to PDF file format, rather than another image file format, was preferred due to the facility that PDF management software offered in converting the text through OCR. This rendered machine-readable, editable text formats (still in PDF) that were then converted into a spreadsheet. Furthermore, the software being used (Adobe XI version 11.0.09) made it easier to batch process and clean these PDF files from granular, digital noise resulting from the scans, as well as applying OCR. A similar process was used to digitise the item analysis presented in the reports. Table 6-1: Example of discrete sets of data collected in a spreadsheet shows an example of discrete sets of data collected in a spreadsheet format following this process.

6.3.2.1 Processing the PDFs

In processing the scanned files, the steps taken required that each file was:

i. cropped and de-skewed to retain the relevant data, remove margin lines from the edges and straighten the PDF image file (The National Archives, 2016)
ii. processed to sharpen the image and remove any blur
iii. de-speckled to clean the pages and remove granular marks that resulted from the scan
iv. converted to monochrome rendering a black-and-white document
v. processed through the OCR facility to create a machine-readable editable text format document (Commonwealth of Australia, 2013)

Once this was complete, the scanned documents could be converted to a spreadsheet and establish the digitised record of the data set.

6.3.2.2 Converting to Excel

The conversion from a PDF document to an MS Excel spreadsheet was done in such a way as to create a single workbook with several worksheets. To this end, 28 PDF documents were converted into 28 workbooks, two for each examination cycle, and each worksheet represented a single page of the source record.

The general process for conversion and verification of the source record to the digitised record required:

i. conversion to a Microsoft Office Excel format (XLSX) to create a workbook of the data with multiple worksheets (one for each scanned page)
ii. first stage error checking procedure to inspect for slippage and broadening. Vertical slippage and horizontal broadening refer to the undesired shifting of data on the output spreadsheet and are explained in further detail in the following sections
iii. collation of worksheets into a single worksheet that represented the entire data set for that examination cycle
iv. second stage error checking for quality and integrity
v. the amalgamation of the two spreadsheet files into a single file thus producing fourteen data sets – one for each examination cycle
vi. final data verification and data set integrity check

As there were two sets of records for each year, this process resulted in the creation of two spreadsheets that together contained a complete record of outcomes for that year's examination cycle. Each pair of files would later, following the error and integrity checking

process, be combined into a single spreadsheet representing the single examination cycle and their respective results.

Table 6-1: Example of discrete sets of data collected in a spreadsheet

Index	School ID	Student ID	Name	Social	Maltese	English	Maths	Religion	Outcomes
7	BB			C	C	C	C	C	Passed
8	BB			A	C	B	A	C	Passed
9	BB			C	C	C	C	B	Passed
10	BB			C	C	C	B	C	Passed
11	BB			C	B	D	C	C	Failed

6.3.3 Challenges and considerations for quality control

It should be noted that as the data set is unique to this particular research, the literature review did not return specific procedural information, but rather more generic quality control matters that needed consideration (Archives New Zealand, 2007; Commonwealth of Australia, 2013; The National Archives, 2016). These considerations were mostly associated with governmental or institutional large-scale digitisation processes. Consequently, the process of quality control and data integrity associated with the JLEE data sets needed to be established specifically for this study and relative to the data structures being generated. The main procedures remained in line with the common quality control consideration stipulated in the large-scale digitisation processes.

Following on from the literature regarding the digitisation procedures applied by the Commonwealth of Australia (2013), a thorough review of the digitised data as a quality control measure was implemented to verify that the data conversion was complete and error-free. This required the establishment of a process that would ascertain all data sets had been converted accurately, validate the output, and identify any errors or loss of data in the conversion. This process would also ascertain that each data set — represented by a single row of data in the records — remained aligned during the processing sequence.

6.3.4 Error checking - ascertaining the quality and integrity of the data set

As stated earlier, the worksheets were ultimately collated into a single data set on one worksheet. Error checking and data verification process were then implemented to inspect

the data sets for quality and integrity. The error-checking process considered the quality of the digitised record and made modifications to remedy any errors in the conversion. The data verification process reviewed and ascertained the integrity of the data confirming that the digitised record was a true copy of the source record.

In determining the quality of the conversion and what types of conversion errors may have occurred, the error-checking process showed that although most of the pages had been converted correctly, some spreadsheets required remediation. This was due to the displacement of data within the spreadsheet, additions of spaces or periods and conversion to a non-alphanumeric character such as parentheses, or Greek characters.

Furthermore, the error-checking process took place in two stages: one before and one after the collation process. This was a pragmatic choice stemming from the types of common errors observed during the conversion and how the data was to be managed to retain quality and integrity.

The first stage of the inspection considered displacement of data elements down (slippage) or across (broadening) the spreadsheet (see Table 6-2 for examples of slippage and broadening). This error occurred on a sheet-by-sheet basis which meant that although one worksheet could have been affected, the subsequent worksheets might not have been. It would have proven more complicated to adjust and realign the displaced data elements in a more expansive spreadsheet.

Once the spreadsheets were collated, the second stage of the inspection looked for OCR conversion errors where one character was interpreted incorrectly by the OCR software and was converted incorrectly or had a space or period character added to the grade level result. The fact that the key data was established on the five grade levels (A - E), abs, x, Passed or Failed meant that this stage could batch process the data sets to find anomalies that did not match these elements specifically. The next sub-sections look at these two processes in more detail.

6.3.4.1 First stage inspection: Slippage and broadening

Before collating the worksheets into one, the individual worksheets were thoroughly inspected for slippage and broadening. The scores and outcomes in each row needed to remain aligned with each other so that together they represented a discrete individual data set. The effort to ascertain this integrity focused first on reviewing the digital data for any inadvertent displacement of the data elements within the spreadsheet. This displacement of elements was observed as either slippage down or broadening across the spreadsheet.

Table 6-2: Examples of possible slippage and broadening of data in the spreadsheet

Index	School ID	Student ID	Name	Social Studies	Maltese	English	Maths	Religion	Outcomes
7	BB			C	C	C	C	C	Passed
8	BB			A	C	B	A	C	Passed
9	BB			C	C	C	C	B	Passed
10	BB			C	C	C	B	C	Passed
11	BB			C	B	D	C	C	Failed

a. Original data set

Index	School ID	Student ID	Name	Social Studies	Maltese	English	Maths	Religion		Outcomes
7	BB			C	C	↓	C	C		Passed
8	BB			A	C	C	A	C	Passed	←
9	BB			C	C	B	C	B		Passed
10	BB			C	C	C	B	C		Passed
11	BB			C	B	C	C	C		Failed
						D				

b. Horizontal broadening and vertical slippage

Index	School ID	Student ID	Name	Social Studies	Maltese	English	Maths	Religion	Outcomes
7	BB			C	C	C	C	C	Passed
8	BB			A	C	B	A	C	Passed
9	BB			C	C	C	C	B	Passed
10	BB			↓	↓	↓	B	C	Passed
				C	C	C			
11	BB			C	B	D	C	C	Failed

c. Vertical slippage for multiple elements

In the case of slippage, the data from a single row shifted into a newly created row, pushing all subsequent data downward. Slippage down a column took place by the inadvertent addition of one or more blank cells in the sequence (Table 6-2b. above). Similarly, the conversion occasionally added entire rows or separated a row of data onto two or three subsequent rows (Table 6-2c). For slippage, the spreadsheets were adjusted by moving the data elements to the appropriate cells and removing the blank rows, realigning the associated data horizontally.

In the case of broadening, the data from one or more cells in a single column were displaced to a newly created column, pushing the remainder of the column data outward thus broadening the spreadsheet by creating more columns. Depending on the quality of the scan, the software conversion would occasionally also insert a blank column or separated the subsequent data in a column across multiple columns, thus spreading the data across the spreadsheet over a much broader range of columns (Table 6-2b shows an example of a single column addition). The displacement brought about by broadening was adjusted by moving the displaced data elements back to their cells in the appropriate column, realigning the associated data elements vertically.

The frequency of such slippage and broadening depended on the quality of the scan. Data on the number of times slippage and broadening took place were not collected, however, in most cases, the conversions were true and complete without any displacement of data elements.

Inspecting the data for these errors while all pages were still collected as individual worksheets meant that each sheet could be scanned at a glance to identify errors. At the same time, the worksheets were also reviewed for loss of singular results from the conversion leaving a blank space instead of a grade result. This was remedied by referring back to the source record and adding the missing data accordingly. The fact that this part of the work was focused on single scanned pages allowed reference to be made to the source records to directly locate the page in question. This also allowed the reviewer to confirm that the error adjustments matched the original alignment of data elements.

6.3.4.2 Collation

Once this first stage inspection had been completed for all 28 workbooks, the individual worksheets were collated into a single worksheet. This meant that an entire dataset for one examination cycle was represented on two spreadsheets in preparation for the second stage inspection.

Although these two spreadsheets could have been combined into a single one at this stage, it was considered better to process them separately for the second stage inspection. This would

make each file more manageable and tracking of other errors quicker as the data set being processed was half the size.

6.3.4.3 Second stage inspection: OCR errors

The second phase of the error-checking process looked at errors in the data resulting from incorrect OCR conversion of the results. The customised process to check and correct these errors during this second phase used batch processing techniques to clean the data. Once the initial batch processing was completed, more specific scanning of details was conducted.

The overall cleaning process required the following steps to be undertaken for each data set (workbook file).

i. Indexing: A column was added to the left of the spreadsheet (column A) and an incremental number sequence starting from 1 was inserted so that each row had a unique identifying number. This would be used to return the spreadsheet to its original sequence after having been sorted according to other criteria.

ii. General error corrections: The entire spreadsheet was processed to remove any space or period characters from any of the cells. These characters were unnecessary to the analysis and would have been erroneously added by the OCR process.

iii. Grade result character correction: Character conversion errors resulted from the OCR software converting some grade results incorrectly. C was occasionally converted to left parenthesis "("; D was occasionally converted to an "O"; B was occasionally converted to "β"; Passed – Failed statements written incorrectly; some characters were converted to non-alphanumeric characters.

    To adjust for these errors, the entire spreadsheet was sorted alphabetically by subject column. This was done once for each of the subjects. Sorting the spreadsheet by column resulted in the grades A - E, Abs and x being grouped down the column and any anomalies were separated to the top or bottom of that column allowing the reviewer to identify any errors and take remedial action.

iv. Outcomes corrections: the Passed – Failed outcomes for each of the rows were occasionally changed by the OCR to include different fonts or characters. To this end, the same procedure was followed as in step (iii), separating anomalies, and allowing them to be remedied to correctly read Passed or Failed.

v. Once the corrections were made, it became necessary to ascertain that all outcomes (Passed or Failed) were recorded in the same way throughout each document to facilitate the counting of identical terms. To this end, the spreadsheet was re-sorted

one more time according to outcomes such that all rows with “Failed” statements were separated from those with “Passed” statements. Subsequently, each of the “Failed” statements was made identical as were the “Passed” statements.

vi. The spreadsheet was finally re-sorted according to the assigned index number in column A and returned to its original sequence.

Any remedial actions to the data sets were done with direct reference to the source document to ascertain that the grade level result remained the same. The School ID, Student ID and Student Name were used as row indexes so that any issues arising from errors could be compared to the original scanned PDFs and error correction implemented. Once this process was completed, the names of the students were no longer relevant to the research and were removed, thus anonymizing the data, and avoiding issues of identification or association with any one individual.

Having completed the process for all six of the sorting criteria (five subjects and pass-fail), a final sorting was done according to the index number allowing the data set to be returned to its original sequence.

6.3.5 Amalgamation and verification

Once the error checks had been completed, the two spreadsheets representing the examination cycle were combined so that the entire data set was amalgamated into a single spreadsheet. This resulted in fourteen spreadsheet files, one for each cycle that had been processed and edited to have a uniform data format for all.

The amalgamation of each of the examination cycles into a single yearly data set allowed for a complete analysis of the data for that cycle and a comparative analysis across cycles. However, before this could be done, the integrity of the data set as a whole had to be conducted to be certain that each data set was complete and true. This process was made more efficient by having a single complete data set for each yearly cycle that could be manipulated and batch processed as a whole.

6.3.5.1 Integrity check - Data verification

The error-checking procedures were followed by an integrity verification check of the processed data set to confirm that the individual data set for every row on the digitised record matched the same set of data in the source record.

Three comparative methods were used to confirm alignment and integrity: interval comparison of digitised and source records; count verification of individual data sets; criteria comparison against outcomes.

i. Interval Comparison

The interval comparison was done at varying intervals along a data set by scanning down the spreadsheet and cross-checking the information of a block of five rows every three hundred rows against the source record to confirm that the scores had remained aligned. Any variation in the block would reflect a misalignment further up which could be traced and corrected.

ii. Count Verification

A second method used was to compare various counts from the digitised record to the official count in the examination report (Educational Assessment Unit, 1997 - 2010). These records had the official total number of applicants who were registered to take the examination for each year. This number and the number of individual data sets on the digitised record would need to be the same. A discrepancy would require that the cause of the discrepancy be traced and remedied.

Furthermore, a count of the number of A's, Bs, Cs, Ds, Es, Abs and x's in each subject column and a comparative of this count to the total number of registrants would allow confirmation that each element in each cell in that subject column is correct. If the count comparison varied, then one or more of the grade results was incorrectly digitised and needed to be traced and corrected.

iii. Criteria comparison

This final step made use of the conjunctive criteria applied by the EAU to double-check outcomes for each row of data. The criteria rules were coded as an MS Excel formula and applied to each row generating a "Passed" or "Failed" result for that row in a new column. The result of the formula was compared to the outcomes on the digitised record and should have been equal for each row. If the two were not equal, then that signified an issue that needed to be traced and remedied. The records were reviewed repeatedly until the criteria comparisons matched the source record for each of the individual data sets.

6.3.6 Schematic Representation of the Digitisation Process

Figure 6-2 Digitisation Process (i)

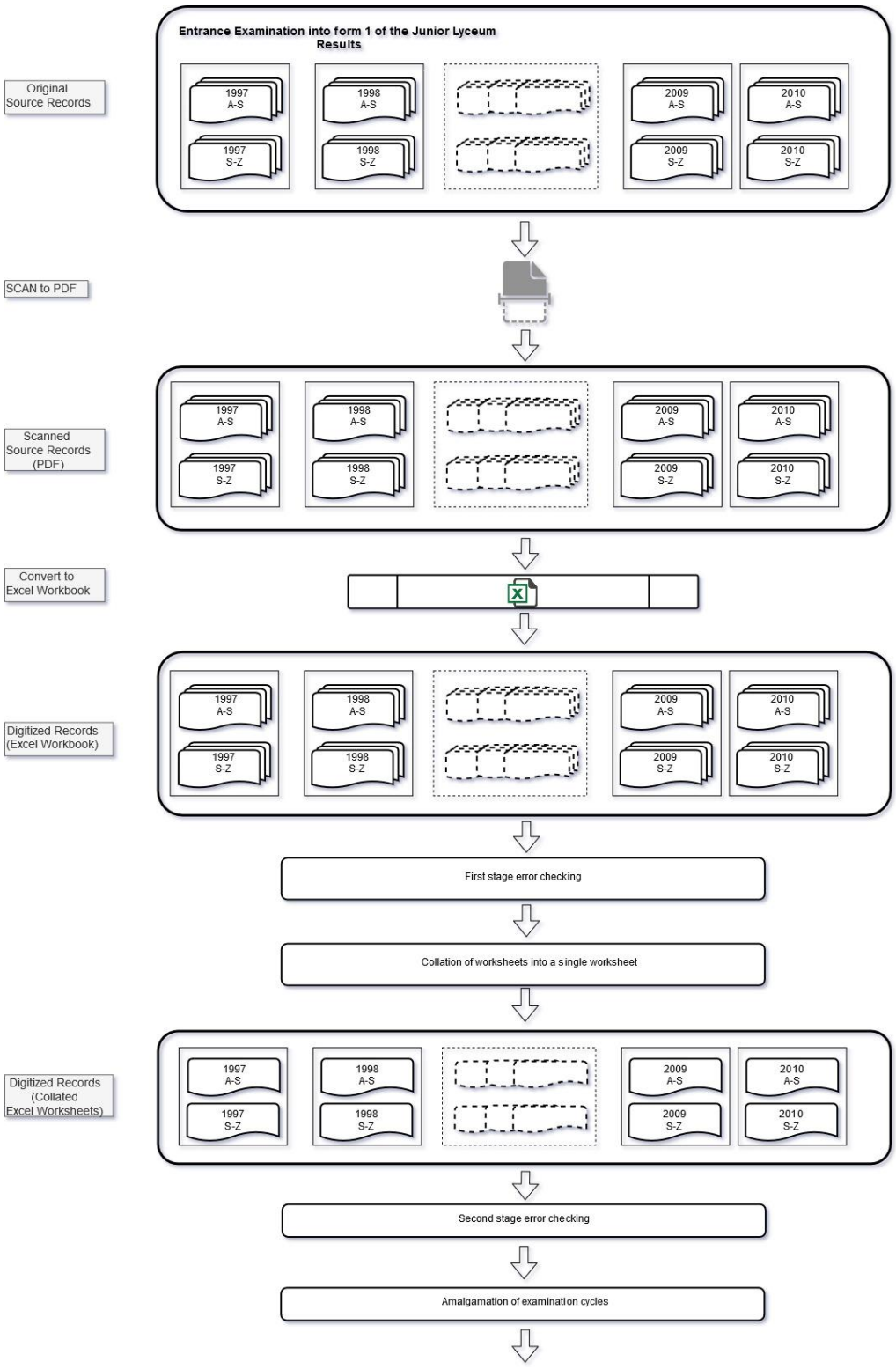
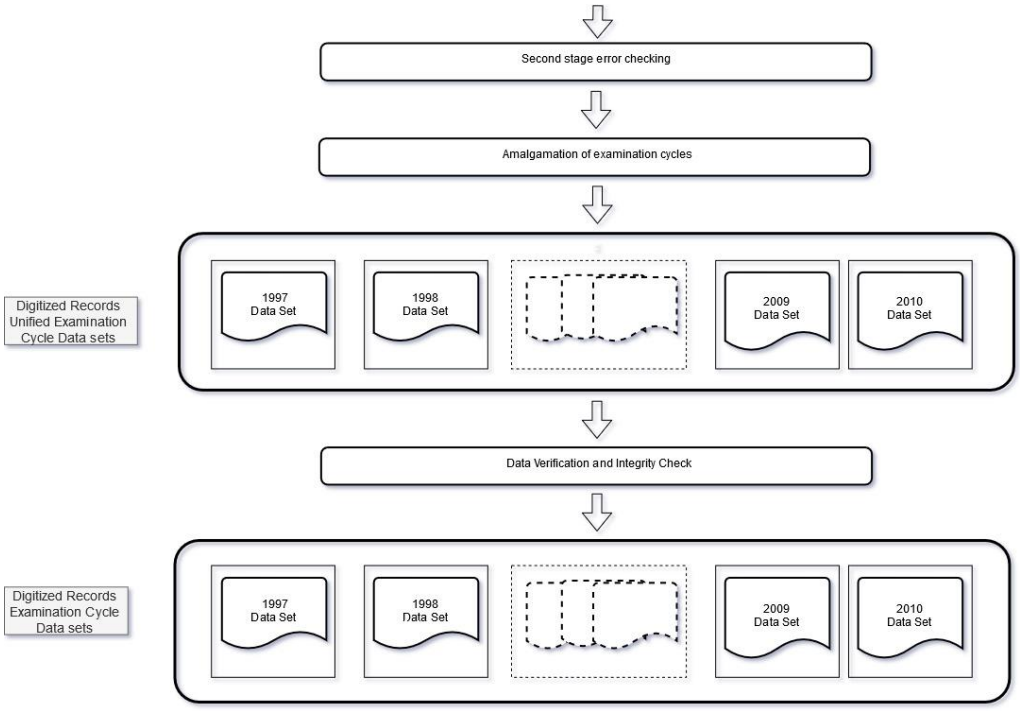


Figure 6-3 Digitisation Process (ii)



6.4 Processing Maltese texts to determine readability

This section outlines the process developed and applied to analyse Maltese comprehension texts to determine readability and determine reading complexity.

As discussed earlier, various online readability tools that could process the Maltese comprehension text using LIX, ARI and CLI proved to be unreliable and inconsistent when their outcomes were compared across platforms. This was not the case when processing the English texts. The analysis therefore required that a dedicated process be developed to analyse Maltese texts using these three readability algorithms.

Each of the scanned Maltese texts (sans title) was parsed using a text splitter (*Split Text - Online Text Tools*, n.d.), producing a list of all words from the text with any associated letters, numbers and punctuation attached as required by the ARI and CLI definitions. The Maltese language is particular, however, in having the definite article that is directly attached to the word using a hyphen (e.g., il-kelma). The definite articles were separated from the word so that each article together with its hyphen constituted a separate word on the list. The list was then transferred to an Excel sheet so that each word created a record for analysis (one word per row).

In preparing the data for readability analysis, the number of characters of each record was first counted using Excel's "LEN" function. This count was added to the record as a new field (Word

Length). Excel tools were then applied to the complete set of records to determine: the total number of words (W); the total number of sentences (S); the number of words with 6 characters or more (X); the total number of characters (L); the average number of characters per word; and the average number of words per sentence. Having determined these measured variables, an algorithm was written using Excel formulae to determine the LIX, ARI and CLI values based on the following equations drawn from Tillman & Hagberg (2014).

$$LIX = \frac{W}{S} + \frac{X \times 100}{W}$$

$$ARI = \frac{4.71L}{W} + \frac{0.5W}{S} - 21.43$$

$$CLI = \frac{5.88L}{W} - \frac{29.6S}{W} - 15.8$$

Where:

W = Total word count
S = Total number of sentences
X = Number of long words with 6 letters or more
L = Total amount of letters, numbers, and punctuation marks

Although ARI and CLI outcomes correspond to the grade levels used in the United States, they were not relevant to determine Maltese grade levels but remained applicable as a comparison tool for the different texts.

On the other hand, the LIX outcomes are scale-based and do not correspond to any grade level. It therefore differs from the ARI and CLI tiers and the outcomes needed to be processed on a different scaled graph. Nevertheless, as all three formulas measure variations in readability, then trend similarities between all three graphs would reflect variations in text complexity.

6.5 Cognitive item demands - Process

The cognitive demand presented by the various test forms was analysed using a process similar to that used by Jones et al. (2009). This established a categorisation framework using Bloom's taxonomy to determine the cognitive demand characteristics of test forms according to question verbs used for each test item. However, this method was superficial and

considered the question verb solely and directly. It did not consider deeper cognitive demands that would be made by multiple process steps that may have been integrated into the question.

The different categories used in Bloom's Taxonomy (Anderson & Bloom, 2001) were aggregated for the analysis as shown in Table 6-3: Cognitive level categories below. This was done to reduce the ambiguities mentioned earlier in determining the intended and received meaning posed by the question verbs associated with remembering and understanding.

Table 6-3: Cognitive level categories

Bloom's Taxonomical Categories	**Aggregated Cognitive Levels**
Remembering (R)	Lower Cognitive Level
Understanding (U)	
Applying (A1)	Intermediate Cognitive level
Analysing (A2)	
Evaluating (E)	Higher Cognitive Level
Creating (C)	

As a result, each question item could be classified as lower, intermediate, or higher cognitive levels and a frequency could then be counted for each category. Furthermore, this was done for intermittent years (1998; 2000; 2005; 2009) for each of the five subjects. The question verb for each test item was recorded on an array as shown in Table 6-4 below.

The data presented in Table 6-4, Table 6-5, Table 6-6 and Figure 6-4 use partial data extracted from the English 1998 data set. The full and complete data set is presented further on.

Table 6-4: English 1998 - question verb array for each question and sub-question

Question	**1**	**2**	**3**
A	underline	underline	underline
B	write	write	write
C	Complete	Complete	Complete
D	conjugate	conjugate	conjugate
E	Underline	Underline	How
F	Explain	Explain	Explain

The array items were then classified against the prepared verb list to assign a cognitive level to each item according to its question verb. To simplify this process, each cognitive level was

assigned a number with “Remembering” being assigned 1 and “Creating” being assigned a 6, and all other categories assigned 2 -5 respectively as shown in Table 6-5 below.

Table 6-5: English 1998 - Question and sub-question categorisation according to cognitive level

Question	1	2	3
A	1	1	1
B	1	1	1
C	3	3	3
D	3	3	3
E	1	1	2
F	4	4	4

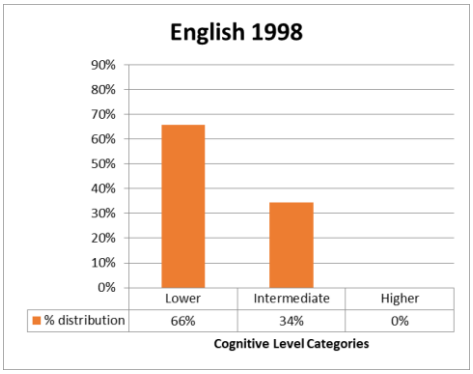
The cognitive levels were then counted, and the percentage distribution was calculated based on those counts. A cognitive profile could then be determined based on the percentage distribution of each category as shown below in Table 6-6.

Table 6-6: English 1998 - Question item distribution according to cognitive level

	Cognitive Levels	Count	%		Aggregated Cognitive levels	Count	%
1998	**R**	21	41%		**Lower**	26	51%
	U	5	10%		**Intermediate**	24	47%
	A1	20	39%		**Higher**	1	2%
	A2	4	8%				
	E	0	0%				
	C	1	2%				

The percentage distribution levels were charted for each of the aggregated cognitive levels as shown in Figure 6-4 below, and compared for each subject to determine longitudinal variations in cognitive levels. This would then reflect on any general shifts in the cognitive demands made by the examinations.

Figure 6-4: Cognitive level profile for English (1998)



6.6 Item response data

The discrimination and facility indices were presented in the EAU reports as a table showing D_i and F for each question and sub-question of each examination. These were only available as hard copies in these reports and needed to be converted to a digital format to run any statistical analysis. These psychometric tables were converted using a process identical to that used to convert the grade scores and described in the previous sections. The tables were scanned to PDFs and transferred onto a spreadsheet. Reliability of the conversion was done by reviewing and comparing the hard and soft copies of each table. Once digitised, the tables were available for statistical analysis and used to plot annual graphs of D_i vs F.

6.7 Summary

The digitisation and preparation of the record of results, item analysis and parts of the examination texts were required to synthesise a manageable set of data that would support the analysis. In doing so, this chapter also responded to the first RQ in establishing mechanisms, tools, and procedures to aggregate and process the JLEE data for analysis.

The main outputs and procedures used for the digitisation applied tried-and-tested methods to convert large volumes of printed material, thus creating a soft record of the data that could be analysed using software. These were applied to the record of results and the item analysis.

The conversion of the English comprehension texts was a straightforward OCR conversion applied before using online readability software, however, the Maltese texts proved to be more challenging to analyse. The procedures associated with analysing Maltese texts for readability required the development of new reliable procedures that could be processed using the LIX, ARI and CLI algorithms. As the available online tools proved unreliable for Maltese, a complete process was devised to parse the passages, adjust the outputs according to the language characteristics, and comparatively analyse the texts.

Section 3: Analysis and results

Section Overview: Analysis of policy, context, and outcomes

This section presents the analysis and respective results from the three sets of records — the two policies (NMC; FACTS), the achievement and results data (1997-2010), and the associated reports (1997 – 2010). Each set of analyses and results is presented in a separate chapter and responds to the research questions using the methodologies described in section 2. The underlying framework for analysis remains the i-p-o-c quality framework described in earlier chapters, and it structures the overall analytical sequence into their respective chapters.

Chapter 7 considers the input and context domains with a focus on the introduced policies. This policy analysis was applied to the NMC and FACTS documents to determine intention and scope and understand influencing factors acting on the educational landscape during that time.

Chapter 8 informs the process and context domains analysing the examination reports to determine variations in test construct as well as fluctuations in difficulty levels and discrimination power of the test forms.

The closing chapter of this section is structured to analyse variations in outcomes and achievement and informs the output domain. This was structured around a time-series analysis of results data and used to determine variations in pass-fail rates over the years which, along with other longitudinal and cross-sectional traits analyses, would respond to the third research question.

All three document sets were initially reviewed and considered for authenticity, credibility, and representativeness against Scott's (1990) criteria after which the analysis process looked for meaning and interpretation.

Authenticity, Credibility, and Representativeness

Scott's (1990, p. 19) criteria for authenticity, credibility, and representativeness were applied to the available document sets to test their validity and reliability and present quality support arguments about the sources of data and information being used as the principal resource (Scott, 1990, p. 36).

Authenticity: All three sets of documents were genuine and of unquestionable origin. The results data and associated reports were collected directly from the EAU's administrative offices and the policy documents from the official government website (*Policy Documentation Archive*, n.d.). Furthermore, all are referenced repeatedly throughout the research and government reporting literature (Attard Tonna & Bugeja, 2016; Bezzina, 2003; Borg & Giordmaina, 2012; Cutajar, 2007; Grima et al. 2008; D. Mifsud, 2015).

Credibility: As all three documents were produced directly by the Institutions responsible for their keeping, they are accepted as evidence drawn from the official record. The procedures and checks described earlier asserted that the digitisation of the data records retained integrity and quality, and produced a true digitised copy of the original data record thus maintaining credibility.

Representativeness: The policy documents collected are the true and complete documents issued by the government at the time. The achievement and results data present the JLEE results for each of the years in question and are a complete record of those students that sat for the examination.

Both the EAU's official documentation and the Ministry's online policy documents satisfied the criteria of authenticity and credibility, and represented all documentation that considers:

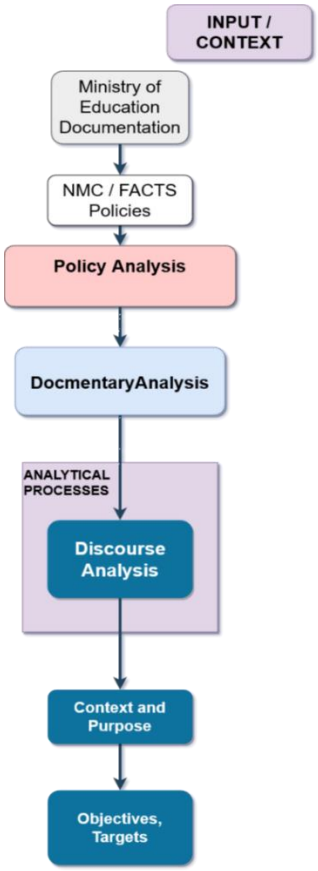
1. The official published educational policies in question,
2. the Junior Lyceum Entrance examination records of achievement for the periods 1997 – 2010, and
3. The Junior Lyceum Entrance Examination reports for the periods 1997 – 2010.

7 Inputs and context: Policy analysis

7.1 Chapter overview

An analysis of the first two policy documents — the NMC and FACTS — informed the input and context part of the analysis determining if there was any particular scope, purpose, or systemic change in context that may have driven or influenced changes in student outcomes.

Figure 7-1 Analysis of input and context flow diagram



This chapter reviews the discourse set out in those policies and allows the deconstruction of the text to identify statements of intention relevant to student outcomes. It also establishes the contextual scenario and sets the stage for interpretation of results from an analysis of the process and output domains. In doing so, it explains the new educational contexts brought about by the policies, identifies purpose and objectives, and determines if any targets were set to specify successful implementation. It also includes an inquiry to determine if outputs or intended outcomes were projected beforehand or set as performance indicators.

Furthermore, the chapter includes a brief analysis of the third policy document — the NCF — published in 2012 to contrast its discourse to its predecessors' and identify similarities and changes in the policy objectives across all three documents.

7.2 NMC and FACTS — the contextual backdrop

The NMC was radical in that it intended to introduce a more liberal approach and allow schools to implement the curriculum in their own way. This presented concepts such as school-based action research (MEYE, 1999, p. 80) and the responsibility to draw up plans for implementing the new minimum curriculum. Bezzina (2003, p. 16) claims that in acting to implement the NMC, the main goals remained that of improving *"the quality of education for all students.*" He also highlights other key changes associated with the intended shifts introduced by the NMC, including a structured move towards more student-centred systems of education (Bezzina, 2003).

FACTS then took the concept of shifting decision-making processes away from a central authority to a different level. It decentralised the educative processes from the central authority to the colleges (regionally designated primary and secondary school clusters under one principal), delivering a major shift in the governance of educational institutions in Malta and Gozo. Although management and leadership responsibilities were effectively transposed, as were most systems of education and schooling, the actions remained underpinned by the policy's intention to achieve improvements in quality education. These intended improvements targeted a broader set of domains highlighted by the "Knowing Our Schools" document: "1. Management, Leadership and Quality Assurance; 2. Curriculum; 3. Learning and Teaching; 4. Attainment; 5. Support for Students; 6. Ethos; 7. Resources" (MEYE, 2004b, p. 7).

Furthermore, different functionality was established within the educational system to support these changes including the restructuring of the education division into two directorates: Directorate for Quality and Standards in Education (DQSE) and the Directorate for Educational Services. Both directorates were a direct result of the FACTS policy as stated in the annual

government department reports (Operations and Programme Implementation Directorate (OPM), 2007, p. 285).

Both policies were therefore far-reaching, ambitious, and intended to work together to transpose educative processes and the local systems of educational governance. If, as intended, the changes did make an improvement in the quality of education, then such a change should be observable in the different domains listed above, including attainment.

7.3 Policy purpose and objectives

Discussions presented earlier argued that an analysis of policy establishes a comparative context (Bowen, 2009) by documenting intentions, purpose, meaning, and targets set by the policy writers (Codd, 1988; Hill & Varone, 2016; Taylor et al. 1997b). The analysis of content of the three policy documents looked to understand their intended effect and establish a clear contextual backdrop against which to interpret the other analytical processes (Codd, 1988; Hill & Varone, 2016, p. 5).

Although the main two policies being considered are the NMC (1999) and the FACTS (2004) policies, the inclusion of the NCF (2012) was purposed to get a further look at intentions by linking continuing or evolving discourse across the three policy introductions. To this end, the review highlights a common theme focused on improving "quality education" in all three documents and features as a principal tenet of policy change in all three (Attard Tonna & Bugeja, 2016; Borg & Giordmaina, 2012; Cutajar, 2007; Grima et al. 2008; Pisani et al. 2010).

The introductory statements in each of the policy documents (Calleja & Grima, 2012; Cristina, 2012; Galea, 1999, 2004; Mizzi, 1999, 2004) established the intention to deliver improvement and positive progress to the Maltese educational system through each of the policies respectively. All three documents associate the concept of improvement and positive development with better quality education, and all three use the term "quality" throughout to refer to the concepts and purposes of improved education. The FACTS policy embeds the term in the title itself: *"A New Network Organisation for Quality Education in Malta"* (MEYE, 2004a).

The repeated reference to the term "quality" throughout the three main policy documents, in association with delivered educative processes and learning outcomes, underscores the continuous intention to strive towards better education for learners even though the discourse may have varied slightly between documents.

> *NMC: "Principle 1: Quality Education for All...The National Minimum Curriculum recognises the right to quality education as the main aim of this process of curricular review." (MEYE, 1999, p. 23)*

FACTS: "The next phase in Malta's education development is to ensure quality education for all." (MEYE, 2004a, p. xix)

NCF: "Every child is entitled to a quality education experience and therefore all learners need to be supported to develop their potential and achieve personal excellence." (MEDE, 2012, p. 32)

As discussed in 2.3.1 Sustaining effective reform, and pointed out by Sammons (2009, p. 123), in the drive to achieve change for the better, policies will explicitly reflect the general intention of decision-makers to deliver reforms that promote educational improvement, quality, and raising of standards. The discussion also pointed out that such large-scale efforts tend to be more effective *"to yield substantial change in proactive and student outcomes"* (Nunnery, 1998, p. 285). The general intention of all three policies was intended to enhance quality across the educational system leading to better education for learners. What was missing from the NMC and FACTS (but not the NCF) were integrated systems for determining the impact of such changes.

7.4 Gauging quality impact

Both the NMC and FACTS policies maintained a broad and general use of the term "*quality*" and associated definitions of *"Quality Education for All"* (Galea, 2004), however, neither of them specified how improved learning, achievement, or outcomes were to be monitored and evaluated. The earlier review of the literature (2.3.1 Sustaining effective reform) emphasised that these efforts need to be grounded in effectiveness research (Borman et al. 2003; Creemers & Kyriakides, 2007), and that link should be explicitly integrated into the policy document or reform action itself. However, none of the literature or official documents presented any such mechanisms or investigations, and the policies themselves seemed to support localised, in-school monitoring systems where schools self-determined if progress in achievement, amongst other factors, was acceptable.

In his introductory message at the beginning of the NMC document, the director-general of education in Malta recognises the need to monitor the implementation of the NMC through research to "objectively gauge" the impact of policy on Maltese education (Mizzi, 1999). However, there are no concrete statements or suggestions of how to ground these reforms in any structured system of effectiveness research. The FACTS policy does not describe any specific measures or considerations to this end either. The intention to monitor and gauge implementation and impact was therefore acknowledged, yet not structured, or tethered to any set of key performance indicators (KPIs) within the policies themselves.

Such standards were not, however, absent from the educational landscape. A separate standardised procedural document was delivered in 2004 establishing a school audit process and measures of quality. The document was published —"Knowing Our Schools" (MEYE, 2004b) — and introduced a School Internal Audit Process creating structures to define KPIs (therein referred to as Quality indicators (QI)) in the different educational domains. The document delivered a standardised manner in which to measure quality education. One of the seven key areas was specific to attainment (MEYE, 2004b, p. 37) and noted that considerations were needed to articulate how students compared to their peers concerning "*benchmarking, the National Mean and/or grades obtained in external examinations."*. This document remained distinct from the policies, however, whilst being indirectly affiliated through the underlying concepts of quality education and the general impetus of the NMC and FACTS.

The NCF, by comparison, does pay more specific attention to such measures and considers year-on-year attainment statistics to be a function of the Quality of Education:

> *"The statistical targets set at the end of compulsory schooling are not an end in themselves. On the contrary, they are to serve as success criteria and achievable goals which we can realistically work towards." (MEDE, 2012, p. x)*

The statistical targets establish a set of KPIs that are to reflect the attainment targets of the new changes in the curriculum framework and are pre-defined in the document: Table 1: Outcomes of Education 2012 and Targets Set for 2027 (MEDE, 2012, p. 24). Similar targets for student learning outcomes, or achievement targets, are not as explicitly obvious anywhere in the document nor are they defined in the previous two policies under consideration here.

7.5 Summary: Inputs and context Analysis

This analysis shows a continuative nature to the three policies in terms of structuring and delivering "quality" education across the Maltese educational landscape. There is also a developmental thread passing through all three documents that exhibit signs of evolving processes, with steps being taken in the subsequent policies to improve development, implementation, and monitoring systems for enhancing the quality of education.

In terms of actual impact measurements, the available literature is scant. Pisani et al. (2010) reviewed the impact of the NMC on equality and Borg & Giordmaina (2012) investigated the impact of the FACTS policy on school and college personnel, both cautiously concluding that perception of impact was generally positive. However, other impact studies or measurements resulting from the NMC and FACTS on other educational systems of teaching and learning were not readily located. Considering the introduction of such broad-scale policies

underpinned by “paradigm-shifting” intentions (a new student-centred curriculum and decentralisation), it would be difficult to argue that there was no impact on schools or students. However, there are no tangible measures of impact that can be used to make such arguments.

In reviewing the input domain of this study, what can be said is that the policies and supporting documents associated with their rollout and support systems do deliver on the intention to change the Maltese educational system. What this implies is that, as argued earlier, this should be reflected in positive changes in student outcomes and attainment and observable in the output domain.

8 Process and context: EAU reports, constructs, and forms

8.1 Chapter Overview

The analysis presented in this chapter responds to the second research question being considered in this study. It informs the process and context domains of the research framework reviewing test constructs for continuity and consistency, and investigating test forms for changes affecting the mental load of the examinations.

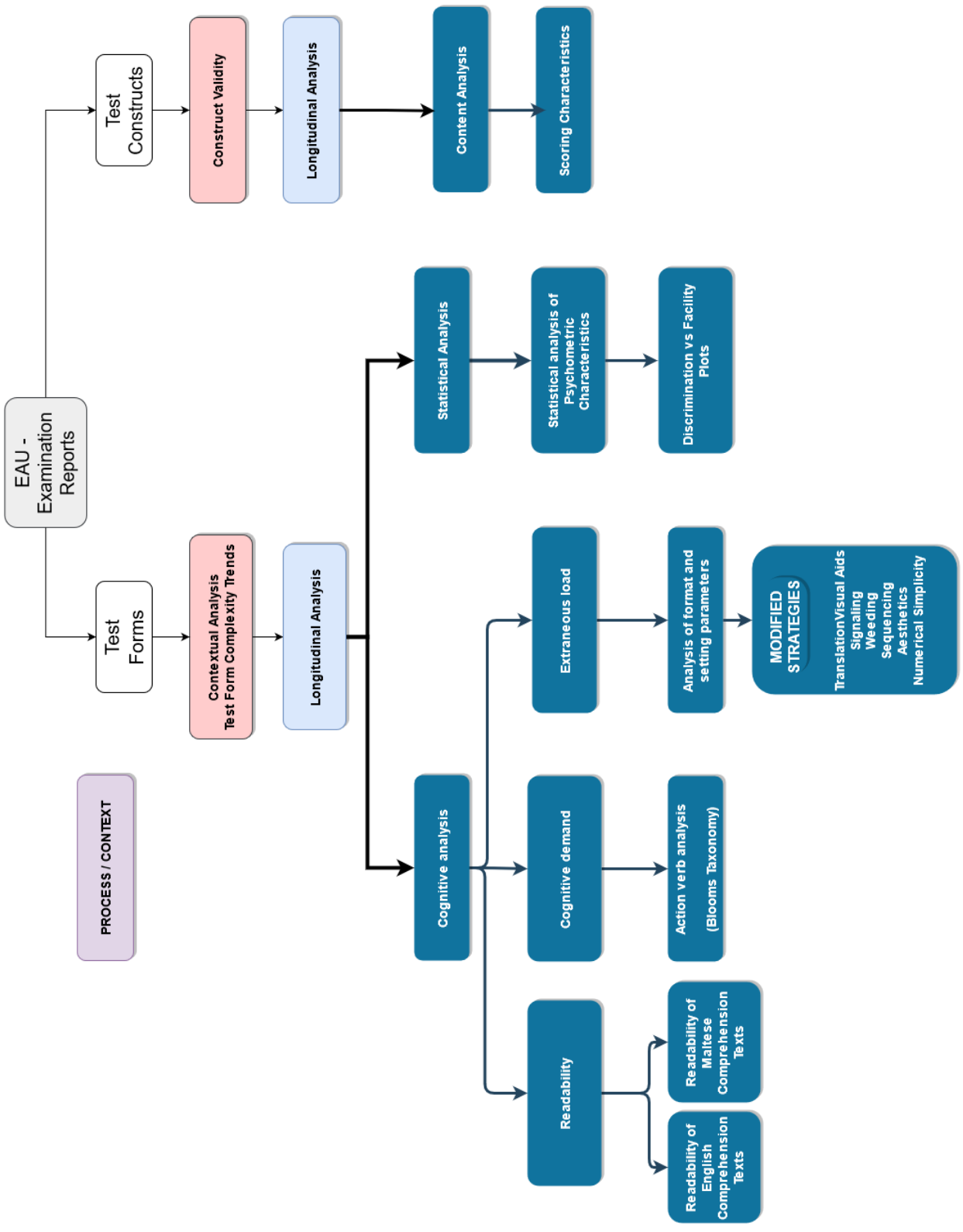
This part of the study proceeds along both these analytical pathways and presents a multilevel analysis of process and context in three parts:

i. Construct analysis: analyses the content and statistical distribution structured in the various specification grids to inform construct continuity and consistency,
ii. Cognitive analysis: reviews test form characteristics that would be considered factors affecting complexity and mental load,
iii. Statistical analysis: analysis of psychometric characteristics from each report to understand any variations in facility and discrimination and possible interpretations.

These analyses inform an understanding of the examinations' difficulty levels and changes to exam quality and standards. This chapter acts alongside the policy analysis to add further perspective to the outcomes analysis by looking at cognitive demands and test item complexities.

The different pathways for analysing these different affecting factors were dependent on trend analysis and are outlined in Figure 8-1.

Figure 8-1 Analysis of process and context flow diagram



8.2 Construct analysis – linking constructs

To inform the construct analysis, the study developed a set of longitudinal tools to compare the main aspects included in the EAU specification grids: i) content specifications, ii) scoring distribution, and iii) anticipated difficulty level distribution. By comparing a sequential array of these construct characteristics, the study was able to both link the constructs across the fourteen years and understand the degree of longitudinal continuity and consistency. This would subsequently inform an analysis of possible changes to the intrinsic cognitive load.

To compare constructs, the three analytical tools processed data from 1998 to 2010. There were, however, a few particularities that need to be noted:

- 1998 was selected as a starting point, rather than 1997, as it recorded all three data aspects in the specification grids, whereas the 1997 report did not.
- Social Studies examinations were discontinued after 2009.
- The 2000 and 2005 data coincided with the introduction of the NMC and FACTS policies.
- Revised examination papers in line with the new syllabi for English, religion, and Maltese were implemented in 2006 (Grima et al. 2008, p. 107), and mathematics followed suit in 2007 (Curriculum Department & Educational Assessment Unit, 2007, p. 51).

This part of the analysis showed that the effect these changes had on the respective specification grids based on new syllabi was only explicitly noticeable for mathematics, suggesting continuity across the fourteen years.

8.2.1 Content analysis — specification grids

The analysis of content was established around a review of the specification grids and looked to compare the overall score weighting distributions in the different domains. Each of the subjects had its own set of domains and subdomains which remained the same throughout the period in question, except for mathematics (Curriculum Department & Educational Assessment Unit, 1998 - 2010). These domains related to content and outcomes and structured the specification grids in such a manner that the grids related *"…outcomes to content and indicates the relative weight to be given to each of the various areas."* (Educational Assessment Unit, n.d., p. 3). However, there was no common framework applied for the different subjects, and the level of detail used in the specification grids varied. There were notable cross-sectional differences with English and mathematics providing more details about scoring distributions

than the other three subjects which employed a more general overview of sectional specifications.

The analysis presented below was structured on determining longitudinal continuity and consistency for the different subjects. It employs an analysis of content and the associated weighted distribution.

8.2.2 Analysis of the subject specification grids

8.2.2.1 Social Studies:

The specification grids for social studies established three principal domains by which to define the test construct: Content, Estimated Difficulty and Learning Outcomes. Each domain was further subdivided into

- Content: Human; Geography; History
- Estimated Difficulty: Easy; Moderate; Difficult
- Outcomes: Knowledge; Understanding; Skills; Attitudes

The score weighting for each domain and sub-domain are tabulated below in Table 8-1 showing those distributions from 1998 – 2009.

Table 8-1 Social Studies specification grid - score weighting distribution

Social Studies	Content Area (% Distribution)			Projected Difficulty Level (% Distribution)		
Year	Human Environment	Geographical Environment	Historical Environment	Easy	Moderate	Difficult
1998	34	34	32	30	34	36
1999	34	34	32	28	36	36
2000	32	34	34	30	34	36
2001	32	36	32	23	51	26
2002	32	34	34	32	38	30
2003	38	34	28	30	30	40
2004	36	28	36	16	57	27
2005	34	34	32	34	35	31
2006	32	34	34	34	32	34
2007	32	36	32	35	32	33
2008	34	33	33	36	32	32
2009	18	39	43	23	65	12

	Learning Outcomes (% Distribution)			
	Knowledge	Understanding	Skills	Attitudes
1998	32	36	24	8
1999	34	32	26	8
2000	27	35	26	12
2001	29	28	33	10
2002	40	32	17	11
2003	20	29	40	11
2004	17	37	24	22
2005	37	32	31	
2006	34	31	35	
2007	40	30	30	
2008	36	32	32	
2009	27	52	21	

Source: Curriculum Department & Educational Assessment Unit, 1998 - 2010

To highlight anomalies in the distribution, the analysis applied a "heat map" tool using MS Excel's conditional formatting and is presented below in Table 8-2 and Table 8-3. A colour scale is also included in each table. The tool automatically determines the highest and lowest values on the whole array and assigns a graded colour scheme automatically based on percentile distribution: the minimum value (darker shade), the 50-percentile mark (white), and the maximum value (darker shade). The larger the standards deviate, the darker the tinge.

To identify anomalies, comparisons were made within the sub-domain along each column. If the constructs retained a similar standard deviation from the overall array average (33.33), the heatmap would only show a slight variation in colour along that column. On the other hand, darker shades would indicate a greater deviation of the construct weighting from the norm.

The heat maps were analysed together as a single tool rather than independently according to domains.

Table 8-2: Social Studies: Heat map of weighted distribution for Content and Difficulty Levels

Social Studies	Content Area (% Distribution)		
Year	Human Environment	Geographical Environment	Historical Environment
1998	34	34	32
1999	34	34	32
2000	32	34	34
2001	32	36	32
2002	32	34	34
2003	38	34	28
2004	36	28	36
2005	34	34	32
2006	32	34	34
2007	32	36	32
2008	34	33	33
2009	18	39	43

Average		σ	2σ	3σ
33.33	0.00	3.75	7.50	11.25
	σ Colour Scale			

Projected Difficulty Level (% Distribution)		
Easy	Moderate	Difficult
30	34	36
28	36	36
30	34	36
23	51	26
32	38	30
30	30	40
16	57	27
34	35	31
34	32	34
35	32	33
36	32	32
23	65	12

Average		σ	2σ	3σ
33.33	0.00	9.48	18.96	28.44
	σ Colour Scale			

Source: Curriculum Department & Educational Assessment Unit, 1998 - 2010

Table 8-3: Social Studies: Heat map of weighted distribution for Learning Outcomes

	Learning Outcomes (% Distribution)		
	Knowledge	Understanding	Skills + Attitudes
1998	32	36	32
1999	34	32	34
2000	27	35	38
2001	29	28	43
2002	40	32	28
2003	20	29	51
2004	17	37	46
2005	37	32	31
2006	34	31	35
2007	40	30	30
2008	36	32	32
2009	27	52	21

Average		σ	2σ	3σ
33.33	0.00	7.35	14.70	22.05
	σ Colour Scale			

Source: Curriculum Department & Educational Assessment Unit, 1998 - 2010

The heat maps indicate slight variation in construct continuity during 2001, 2003, and 2004, before returning to more consistent patterns from 2005-2008. These variations identified a shifting emphasis from in 2003 and 2004 from the "historical environment" to the "human environment and more importance on Skills + Attitudes rather than Knowledge. There were also slight changes in anticipated difficulty levels in 2001 and 2004.

Otherwise, up to 2008, the distribution remained fairly similar for score weighting according to all three domains, reflecting good levels of construct continuity and consistency over that period. The sudden change took place in 2009 when weighting emphasis was shifted towards the historical environment type questions and the percentage of "difficult" question types dropped by 63% on the average of the previous years.

In summary, the specification grid analysis for social studies therefore shows that within the reference framework defined by the given syllabus, the construct remained relatively continuous and consistent throughout. The 2009 paper was the only exception which, although drawn from the same syllabus, had a notably different distribution matrix compared to previous years.

8.2.2.2 Maltese:

The Maltese examiners constructed the exam around twenty-one learning outcomes (Figure 8-2) that linked the assessment construct to the syllabus. The learning outcome statements (LOS) were organised in two principal domains: Reading, Writing and Grammar; and Writing – Composition. These remained the same from 1998 to 2010.

Figure 8-2: Maltese specification grid - learning outcomes

Learning Outcomes: Maltese Reading, Writing and Grammar	Writing - Composition
The ability to follow the storyline or to find information from the selected text.	The ability to write a composition in a logical, sequential and cohesive way.
The ability to understand vocabulary and the idiom and idiomatic aspects of the language in context.	The ability to plan.
The ability to recognize key relationships between lexical units, syntax and paragraphs	The ability to write without errors.
The ability to spell well.	The ability to write relevantly.
The ability to use good grammar.	The ability to write creatively.
The ability to know and understand the basic terminology of grammar.	The ability to write functionally
The ability to derive basic words from morphological roots of Semitic, Romance and English elements and vice-versa.	The ability to effectively apply Maltese idioms.
The ability to write simple sentences.	The ability to have a sense of audience.
The ability to write compound / complex sentences.	Ability to have good calligraphy and presentation.
The ability to make deductions and inferences as well as interpret pictures/illustrations.	
The ability to use punctuation marks	
Ability to be aware of basic language knowledge.	

Source: Curriculum Department & Educational Assessment Unit, 1998 - 2010

There was no statistical summary on the specification grid, nor any weighting distribution analysis, so a comparative analysis was not possible in this regard. What can be confirmed from the reports is that the construct statements remained constant and did not change over the 13 years being considered.

In summary, the learning outcomes can be confirmed to have been continuous from 1998 to 2010. It cannot, however, be concluded that the construct for Maltese remained consistent, as the weighted distribution on particular learning outcomes cannot be determined due to a lack of detail on the specification grids.

8.2.2.3 English:

The specification grids for English established a set of twenty LOSs for assessing students in line with the syllabus (Curriculum Department & Educational Assessment Unit, 2006, p. 41), and organised into three overarching domains: Language and Grammar; Comprehension (Reading and Understanding); and Composition (Writing) — (Figure 8-3). The specification grids remained continuous for all sittings with one exception that introduced three new statements dealing with comprehension in 2000 (Identify main ideas; Follow a sequence; Recognise the writer's purpose or attitude).

Figure 8-3: English specification grid - learning outcomes

Specification Grid Statements

1.	Candidates demonstrate ability to use	2.	Candidates demonstrate ability to	3	Candidates demonstrate ability to
1.1	Parts of Speech	2.1	Deduce meaning from context	3.1	Plan work (L)
1.2	Punctuation	2.2	Locate specific information	3.2	Write in a logical, cohesive and sequential manner (L)
1.3	Tenses	2.3	Infer from context		
1.4	Phrasal Verbs within a context	2.4	Show awareness of cohesive (connecting) devices	3.3	Write relevantly (M)
1.5	Question Tags	2.5	Understand referring words	3.4	Write functionally (M)
1.6	Direct and Indirect Speech	2.6	Identify main ideas	3.5	Write accurately (H)
1.7	Word order and Sentence Structures (Simple, Compound and Complex)	2.7	Follow a sequence	3.6	Write creatively (H)
		2.8	Recognise the writer's purpose or attitude		

Legend: ✓ = primary objective; * = secondary objective
Language: L = 11% M = 21% H = 8%
Comprehension: L = 8.5% M = 11.5% H = 10%
Composition: L = 10% M = 10% H = 10%
TOTAL: L = 29.5% M = 42.5% H = 28%

Source: Curriculum Department & Educational Assessment Unit, 2005

Although these specifications established continuity through the LOS, consistency in application of the constructs showed variations. The analysis that follows looks at three key characteristics of the construct definition to understand these variations: planning grids, difficulty levels, and marking schemes.

i. Planning grids

Based on the LOS, the assessment developers used a planning grid to map the distribution of learning outcomes over the test sections and respective questions as shown in Figure 8-4.

Figure 8-4: Sample of English Planning Grid 2000

Statement		1.1	1.2	1.3	1.4	1.5	1.6	1.7	2.1	2.2	2.3	2.4	2.5	2.6	2.7	2.8	3.1	3.2	3.3	3.4	3.5	3.6
	Language & Grammar (Writing) Total: 40 marks								Comprehension (Reading & Understanding) Total: 30 marks								Composition (Writing) Total: 30 marks					
Insert: L = low, or M = medium, or H = high against the question number in line with the column statement area being hit.																						
Section I																						
Ex. A	Qs																					
	1	L	✓																			
1 MARK	2	L			✓																	

Source: Curriculum Department & Educational Assessment Unit, 1998 - 2010

The domains remained the same over the 13 years, however, there were changes to the planning grid structures from 1998 to 2001. These grids became more detailed over this period linking the LOS to test items and associating difficulty levels (Low-Medium-High). Starting in 2001 and leading up through to 2010, the distribution planning grid went to a sub-question level and introduced a more granular system that recorded primary and secondary LOS as objectives. This supported better validity and reliability in comparing the subsequent grids for this period.

A longitudinal comparison of the specification grids identified distinct changes in the respective weighting distributions for each of the examinations as shown in Table 8-4 below. From 2001 to 2005, there was a shifting emphasis in the total allocated weighting for the domains. This suggests different priorities being set for different years, and subsequently, a variation in construct due to changing statistical weighting. From 2005 to 2010, the total allocated weighting was more consistent with the scoring distribution becoming more analogous.

Table 8-4: English: % distribution of weightings by domain and estimated difficulty level

	Language				Comprehension				Composition
Estimated Difficulty	Low	Medium	High	Total	Low	Medium	High	Total	L – M - H
1998	*	*	*		*	*	*		*
1999	*	*	*	40	*	*	*	30	*
2000	*	*	*	40	*	*	*	30	*
2001	16.8	15.2	8	40	10.5	12	7.5	30	10-10-10
2002	18	19	8	40	11	8	6	30	10-10-10
2003	17	17	11	45	7	10	8	25	10-10-10
2004	11	14	9	34	9	15	12	36	10-10-10
2005	11	21	8	40	8.5	11.5	10	30	10-10-10
2006	10	21	9	40	4	20	6	30	10-10-10
2007	10	18	12	40	4	20	6	30	10-10-10
2008	10	22	8	40	8	11	10	30[5]	10-10-10
2009	10	19	11	40	4	19	7	30	10-10-10
2010	7	23	10	40	7	16	7	30	10-10-10

L – Low; M – Moderate; H – High

Source: Curriculum Department & Educational Assessment Unit, 1998 - 2010

ii. Difficulty levels

The EAU reports included the proportional weighted distribution by anticipated difficulty level for each of the domains from 2001 – 2010 (Table 8-4). The table shows that there were variations throughout most years in the difficulty level score with 2003 and 2004 being particular as the shift in total scoring distribution affected the anticipated difficulty levels to a greater extent.

Cross-referencing the information in Table 8-4 with the specification grid in Figure 8-3 indicates that although the learning outcomes remained the same over the years, the weighted emphasis varied from 2001 to 2005. The variations in 2003 and 2004 impacted the construct consistency by shifting the statistical specifications between domains as highlighted in Table 8-4 above. However, from 2005 – 2010, the construct became more consistent in emphasis and weighted distribution.

Closer inspection of the reported data showed that the 2002 and 2003 data were erroneously interchanged in the reports. There was a mismatch between the distribution of weighted score presented in the planning grids (Table 8-4) and their respective marking schemes for 2002 and 2003 (Table 8-5). The marking schemes showed that the total language score for 2002 was actually 45 while that for 2003 was 40, with the score for the comprehension section

[5] The weighting distribution is reproduced here as reported in the EAU report 2008 which, for the comprehension, erroneously add up to 29 rather than 30.

changing accordingly. This was important as the analysis of outcomes discussed in the next chapter indicated that the pass rate for English dropped in both 2002 and 2004 when these particular constructs varied.

iii. Marking schemes

A deeper review of the marking schemes for English from 1997 - 2010 showed that there was regular restructuring of the exam construct through broad redistribution of question items and variation in score weightings. These affected the long-term consistency of the test forms, and the effects are tabulated in the two tables that follow.

Table 8-5 shows the year-on-year variations as a weighted distribution of marks according to domain. The table shows distinct changes in the scoring patterns in 1997, 2002, and 2004 with a more consistent distribution for the other years. Although these represent three individual years, an argument for longitudinal consistency with intermittent deviations could not be made once the sections were analysed in more detail.

Table 8-5 Allocated section marks for English test forms

Year / Section	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Lang. & Gram.	45	40	40	40	40	45	40	34	40	40	40	40	40	40
Comprehension	25	30	30	30	30	25	30	24	30	30	30	30	30	30
Composition	30	30	30	30	30	30	30	42	30	30	30	30	30	30
Total Score	100	100	100	100	100	100	100	100	100	100	100	100	100	100

Source: Curriculum Department & Educational Assessment Unit, 1998 - 2010

Table 8-6 below is a review of the statistical changes that took place to the language and grammar sub-sections of the English test forms from 1997 – 2010. During this period, the number of question items used in this section ranged from 30 to 50 affecting the average scoring per item, which averaged between 0.8% to 1.3%. These regular shifts were enough to have had a direct effect on consistency and consequently on the comparative validity of subsequent constructs impacting the intrinsic cognitive load of the test forms as argued earlier (3.5.3.1 Relevant intrinsic factors).

Table 8-6: Change in the number of items and score - language & grammar sub-sections

	Year													
Language and Grammar Sub-Sections	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
No. of Sections	5	4	4	4	5	5	5	4	5	5	4	5	5	5
No. of Question items	36	35	35	40	50	40	40	30	45	40	40	40	50	40
Average Score per Item	1.3	1.1	1.1	1	0.8	1.1	1	1.1	0.9	1	1	1	0.8	1

Source: Curriculum Department & Educational Assessment Unit, 1998 - 2010

Arguments presented earlier stated that there needs to be a consistent statistical distribution for the test constructs as a condition for parallelism (Angoff, 1984; Dorans et al. 2007; Feuer et al. 1998). This condition cannot be deemed true for these English exams implying longitudinal variance in the measuring instrument (Liu & Dorans, 2016) and impacting construct consistency, the overall difficulty levels of subsequent test forms and consequently effecting intrinsic cognitive load. Any score-based inferences cannot, therefore, be equally valid (Pommerich, 2016).

In summary, the analysis showed that although the learning outcomes for the English exams remained continuous from 1998 to 2010, there were different changes to the assessments' statistical characteristics reducing the degree of parallelism between subsequent test forms. The English examination did not therefore maintain longitudinal continuity or consistency from 1998 to 2010. Even during the later period 2005-2010, although there was better congruence between constructs, there were still variations in the mark distribution for the language and grammar sections that affected construct consistency.

8.2.2.4 Mathematics:

The mathematics specification grids were also meticulous in their recording of weighted distributions by question and sub-question. These were organised across two primary grids as stipulated by the guidelines for paper setters (Educational Assessment Unit, n.d., p. 3). Figure 8-5 shows both grids side-by-side — the content distribution grid (left) displayed each question score by subject matter, and the learning outcomes grid (right) identified LOs for each test item. Each was organised into sub-domains which in turn linked to the specific content on the syllabus (denoted by the numbers in the second row Figure 8-5).

The learning outcomes showed the characteristics for each sub-question in terms of Knowledge and Understanding; Skills and Process; and Mathematical Language. A review of

this part of the grid did not lead to any statistical analysis but showed consistency in these characteristics over the entire period 1998 – 2010.

Figure 8-5: Sample of mathematics planning grid 2005

	Problem Solving 20% ± 1%	Number 40% ± 3%							Measurement 15% ± 3%						Shape 15% ± 3%				Data 10 ±1%		Level	K/U	Skills/ Proc	Math. Lang.
Qu. No	1	2.1	2.2	2.3	2.4	2.5	2.6	2.7	3.1	3.2	3.3	3.4	3.5	3.6	4.1	4.2	4.3	4.4	5.1	5.2				
1 a			1																		L	✓	✓	✓
b		1																			L	✓		✓
c																			2		L, L	✓	✓	
2 a																	1				L	✓	✓	
b i					1																L	✓	✓	

Source: Curriculum Department & Educational Assessment Unit, 1998 - 2010

Longitudinal comparisons of consecutive content distribution grids showed that there was a categorisation change to the specifications and weighting allocation in 2007 and the introduction of algebra. The assessment of algebra has been argued by Herscovics & Linchevski (1994) and Knuth et al. (2005) to put more cognitive demand on Year 6 students. At the same time, more emphasis was put on problem-solving while shapes and measurement were merged as shown in Table 8-7 below. There was therefore a change in the construct starting in 2007 that coincides with the introduction of a *"new Mathematics syllabus and new textbook "Abacus"."*(Curriculum Department & Educational Assessment Unit, 2007, p. 51). With these changes in mind, the analysis looked for continuity and consistency in the two separate construct sequences (1998 – 2006) & (2007 – 2010). An analysis of these grids showed continuity and consistency in the applied constructs for each of the two periods as indicated in Table 8-7.

Table 8-7 Mathematics: change in content and weight distribution after 2007

Period of Implementation	Content Area				
1998-2006	Problem Solving	Number	Measurement	Shape	Data
	20%	40%	15%	15%	10%

2007-2010	Problem Solving	Number & Algebra	Shape, Space and Measurement	Data
	35%	30%	30%	5%

Source: Curriculum Department & Educational Assessment Unit, 1998 - 2010

Another source of information drawn from the reports was the anticipated level of difficulty included for each question item. This was part of the planning process so that different levels could be included for different aptitudes (Curriculum Department & Educational Assessment Unit, 2005, p. 50). The reports showed that the weighted distribution by difficulty level

remained similar for all the exams in both construct sets. The heat map below in Table 8-8 shows this consistent longitudinal distribution with an anomalous variation observed in 1999.

Table 8-8: Mathematics – Heat map: distribution of weighting by anticipated difficulty level

	'98	'99	'00	'01	'02	'03	'04	'05	'06	'07	'08	'09	'10
Low	27	21	28	28	26	29	27	27	27	27	27	27	28
Medium	41	42	46	42	43	42	42	42	41	44	43	45	44
High	32	37	26	30	31	28	31	31	32	29	30	28	28

Average		σ	2σ	3σ
33.31	0.00	7.27	14.54	21.81
σ Colour Scale				

Source: Curriculum Department & Educational Assessment Unit, 1998 - 2010

In summary the mathematics examination constructs retained continuity and consistency from 1998 to 2010 with a planned change in construct starting in 2007 and linked directly to the change in syllabus.

8.2.2.5 Religion:

The specification grids for religion were less structured compared to the other reports and offered less information to support a detailed understanding of construct continuity and consistency. Specific links between the test questions and the syllabus were conducted for the 1998 – 2000 papers but seem to have been dropped from subsequent reports. From 2001 to 2010, there are no sections in any of the EAU reports that link the questions on the religion exam to its respective sections on the syllabus.

Table 8-9: Religion – Heat map: distribution of weightings by learning outcomes being tested

Religion	Knowledge	Understanding	Application/Skill
1998	29	30	41
1999	27	32	41
2000	38	31	31
2001	34	32	34
2002	33	34	33
2003	33	29	38
2004	37	30	33
2005	34	33	33
2006	38	26	36
2007	33	34	33
2008	38	38	24
2009	36	37	27
2010	31	33	36

Average	3σ	2σ	σ		σ	2σ	3σ
33.33	-11.71	-7.80	-3.90	0.00	3.90	7.80	11.71
	σ Colour Scale						

Source: Curriculum Department & Educational Assessment Unit, 1998 - 2010

The purpose of the specification grids was to present the distribution of weighted scores pertaining to three main outcome domains: Knowledge, Understanding, and Application in everyday life (Table 8-9). In the 1998 and 1999 reports, the latter domain was divided: Application and Religious Skill.

Table 8-9 shows the changes in weighted distribution according to those domains. An analysis of this table indicates that the variations concerning the selected outcomes were somewhat sporadic and not informed by any specific design that might reflect an underlying scope in the distribution. Except for 1998 and 1999 — which had a slightly different construct because of the application and skills section being separate — the weighted distributions varied within a ±7% range.

In summary, although the examination questions were tied to the syllabus, as stated in each of the EAU reports, the information available was not sufficient to determine if the construct for religion remained continuous or consistent over the period in question (1998 – 2010).

8.2.3 Overview: Linking constructs

The construct creation for all five subjects was structured around the active syllabi at the time. Although Grima et al. (2008, p. 107) stated that from 2006, the English, Maltese, and religion exams were based on updated syllabi published in 2005, there are no statements in the reports that recognise this fact except for the English report in 2010.

In terms of continuity: the construct for social studies, Maltese, English, and mathematics can be confirmed to have been relatively continuous with the noted change in the mathematics syllabus in 2007 and an unusual statistical variation in 2002 and 2004 for English, where score weightings were shifted between language, comprehension, and composition sections. The religion reports did not have enough information to make a definitive determination regarding continuity.

In terms of consistency: The mathematics constructs can be confirmed to have been consistent for both reported constructs on either side of the syllabus change. Social Studies can be confirmed to have remained relatively consistent with some slight fluctuations. The Maltese, English, and religion examinations cannot be confirmed to have maintained consistency in the construct over the years. This is reflected by changes in the statistical distribution and characteristics for English, and the lack of appropriate information for making such a determination for Maltese and religion.

Table 8-10 Summary table of construct continuity

Linking Constructs

Subject	Continuity (1997 – 2010)	Consistency
Social Studies	Yes: Continuity Maintained	Yes: Minor Fluctuations (inconsequential)
Maltese	Yes: Continuity Maintained	No Information available
English	Yes: Continuity Maintained for Learning outcomes. No: Average Score distributions varied	No: Construct changes in 2002 and 2004 only (score redistribution). Consistent across all other years
Mathematics	Yes: Continuity Maintained (1997 – 2006) Syllabus change (2007 – 2010)	Yes
Religion	Not conclusive	No Information available

8.3 Cognitive Analysis — trends in test form complexity

In order to determine the level of similarity, the analysis needed to investigate the cognitive loads for subsequent test forms and understand any degree of longitudinal variation in complexity. The different test forms were established on the same syllabi and intended to retain similar, if not consistent, test specifications (Grima et al. 2008) in terms of extraneous cognitive load. Statistical equating mechanisms could not however be applied due to the nature of the data, and it therefore became necessary to investigate possible changes through different analytical methods that could inform interpretation of any variations in achievement. The discussion presented earlier in section 3.4.4 posited that a difficulty-level change vector may be determined over longer periods to understand particular variations in complexity associated with parallel test constructs. Comparing extraneous factors would reflect on those variations and inform an understanding of such a vector.

In order to determine if there were any overall changes to the difficulty levels, each of the five subject examinations was scrutinised and compared using a framework similar to that applied by Kettler et al. (2009). This was discussed earlier in section 5.3.3 and is represented as part of the schematic Figure 8-1 Analysis of process and context flow diagram. The three main analytical streams discussed in the following sections considered: readability of comprehension texts, cognitive item demands, and the extraneous load associated with the format and structures of the test forms.

The English and Maltese papers alone were compared longitudinally for the readability of the comprehension sections, as it was possible to process these using online algorithms to parse

the text. This was done for successive years. Analysis of cognitive level variations reflected in question-verb changes and general format variations of the test forms could not however be analysed using algorithms, and needed to be processed manually. Both these processes were applied to all subjects in intermittent years —for 1998, 2000, 2005 and 2009 — rather than successive years. In identifying intermittent variations, the analysis was then able to trace those specific changes backwards to determine when the change occurred. The choice of intermittent years was discussed earlier in the methods section 5.3.3.2.

8.3.1 Readability of English texts

The analysis presented in this sub-section is a longitudinal comparison of the readability of the English comprehension texts used between 1997 and 2010. These were processed using the six different readability algorithms as discussed earlier in section 5.3.3.1. The algorithms returned the grade-level readability score recorded in Table 8-11 below. Initial expectations were that each of the different algorithms would produce similar, if not equal, trend patterns for the same texts. Following the initial comparative analysis, as each of the algorithms had a common outcome measure (Laird & Mosteller, 1990) based on a grade-level scale, the data was synthesised into a single average value of outcomes for each year.

Analysis

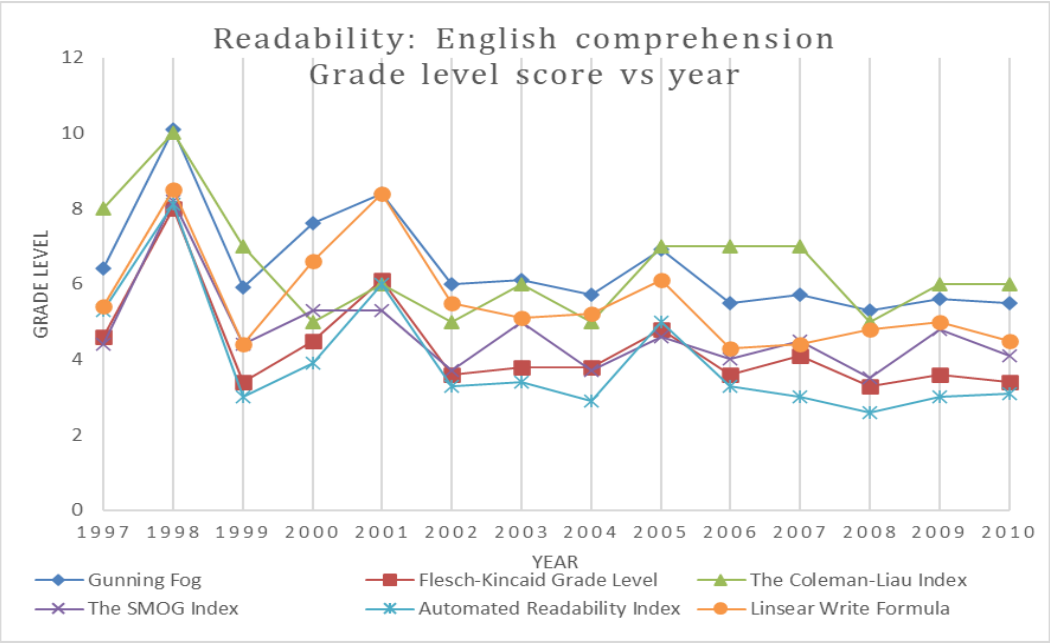
The readability scores are shown in Table 8-11 below. The reading grade levels for each of the readability formula are then plotted against the year (Figure 8-6) and establish a comparative plot of variations in the difficulty level of the texts over time.

Table 8-11: Annual Readability Scores - English Comprehension Text

	Gunning Fog	Flesch-Kincaid Grade Level	The Coleman-Liau Index	The SMOG Index	Automated Readability Index	Linsear Write Formula	Average Readability Score
1997	6.4	4.6	8	4.4	5.3	5.4	5.7
1998	10.1	8.0	10	8.2	8.1	8.5	8.8
1999	5.9	3.4	7	4.4	3.0	4.4	4.7
2000	7.6	4.5	5	5.3	3.9	6.6	5.5
2001	8.4	6.1	6	5.3	6.0	8.4	6.7
2002	6.0	3.6	5	3.7	3.3	5.5	4.5
2003	6.1	3.8	6	5.0	3.4	5.1	4.9
2004	5.7	3.8	5	3.7	2.9	5.2	4.4
2005	6.9	4.8	7	4.6	5.0	6.1	5.7
2006	5.5	3.6	7	4.0	3.3	4.3	4.6
2007	5.7	4.1	7	4.5	3.0	4.4	4.8
2008	5.3	3.3	5	3.5	2.6	4.8	4.1
2009	5.6	3.6	6	4.8	3.0	5.0	4.7
2010	5.5	3.4	6	4.1	3.1	4.5	4.4

It should be noted at this stage that the positioning of each of the plots with the Gunning FOG and CLI holding a higher grade level position and the ARI and Flesch Kincaid grade level maintaining a lower trendline on the graph matches work done by Zhou et al. (2017), who showed similar variations between the readability algorithms. This would likely be the result of the systematic characteristics of the formulae used by the algorithms.

Figure 8-6: English Comprehension Text Readability vs. Year

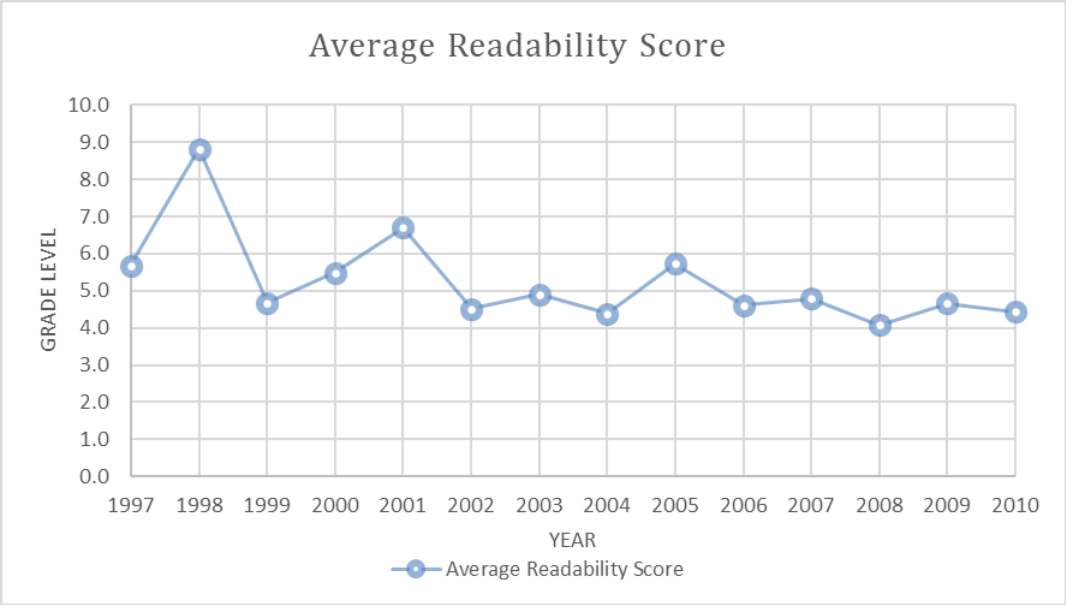


The six plots presented in Figure 8-6 above show that each of the readability algorithms returned distinctly similar trend patterns for the successive English comprehension tests delivered from 1997 – 2010. As has been argued earlier, this reflects on the difficulty level of the text for that year and the general trend over the 14 years.

From the trend patterns shown in Figure 8-6, there are two distinct rises in difficulty levels that are common for all algorithms in 1998 and 2001, with most all other years being otherwise within 1.5-grade levels of each other. 2005 shows another jump relative to the adjacent years before the pattern returns to the same moderate level.

Following the initial analysis of all six readability scores, an analysis of the single average readability score was carried out, plotted in Figure 8-7 below. The plot of combined grade level averages uses the aggregate outcome for the six formulas into a single statistical average (Table 8-11), and thus presents a single difficulty level score for each year. In aggregating the results, the plot of the single average value presents a clearer pattern for interpretation making the longitudinal variations in readability levels more discernible. It also serves to compensate for inherent biases in each of the algorithms used (Levenson et al. 2000).

Figure 8-7: English Comprehension Text Readability - Combined Grade Level Averages



In summary, excluding the 1998 and 2001 readings, the grade level range for the remaining 12 years runs from 4.1- 5.7 with a standard deviation of 0.51. The overall difficulty level for the English comprehension texts therefore remained fairly constant throughout the period in question except for 1998 and 2001.

8.3.2 Readability of Maltese texts

This sub-section examines the readability of the Maltese comprehension texts with the same comparative purpose as the previous sub-section for English texts. The Maltese comprehension texts on each of the papers (1997 – 2010) were processed in the same manner as the English comprehension texts to produce a digitised copy of the text. The analysis required that the readability of the texts be determined using ARI, CLI, and LIX algorithms designed specifically for that purpose (Section 6.4 Processing Maltese texts to determine readability).

Analysis

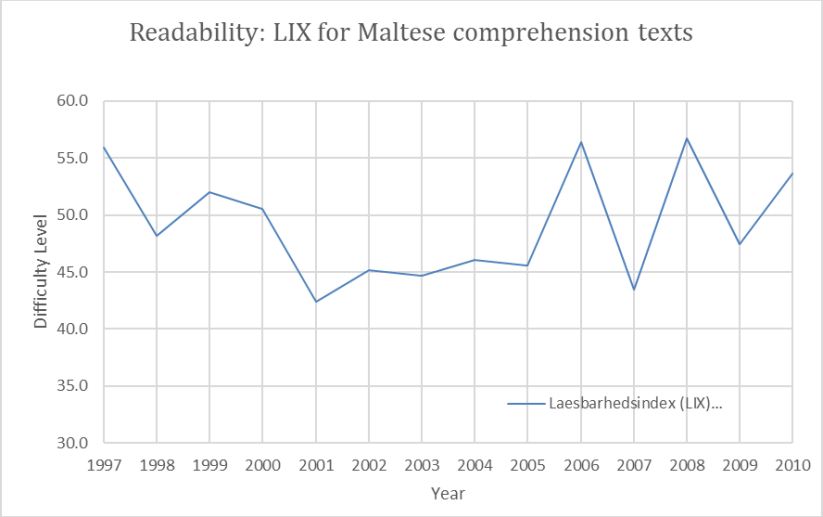
The LIX, ARI, and CLI values for each year are shown in Table 8-12 below.

Table 8-12: Annual Readability Scores - Maltese Comprehension text

	ARI	CLI	LIX
Year	Grade level	Grade level	Difficulty Level
1997	11.1	12.8	55.9
1998	8.6	10.9	48.2
1999	6.6	10.7	52.0
2000	9.0	12.5	50.5
2001	6.2	10.1	42.4
2002	6.0	9.7	45.1
2003	5.9	9.3	44.6
2004	6.1	10.1	46.1
2005	5.9	9.7	45.6
2006	10.4	11.8	56.4
2007	6.1	5.9	43.5
2008	13.5	11.0	56.7
2009	8.0	10.3	47.5
2010	9.0	11.4	53.6

The difficulty levels determined using the LIX formula were plotted against year to establish a longitudinal plot of variations (Figure 8-8).

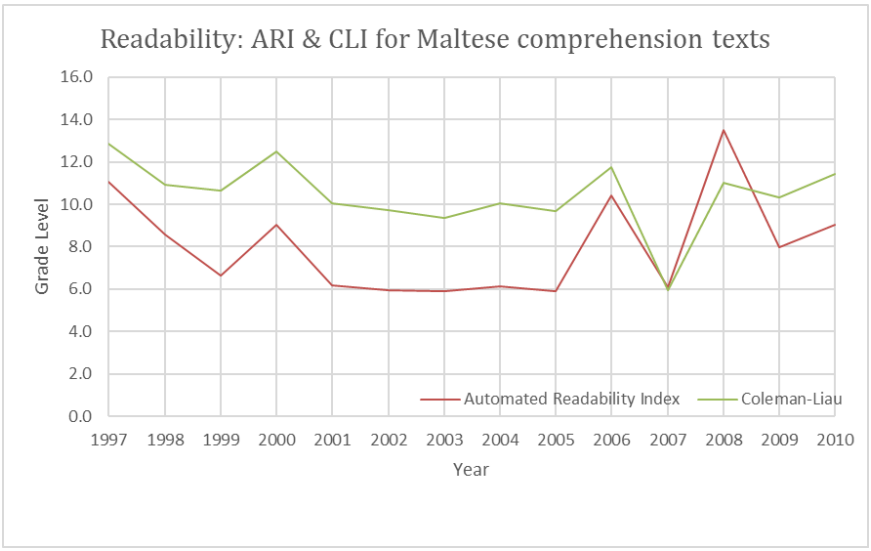
Figure 8-8: Text Readability (LIX) vs. Year



A similar plot was made for the ARI and CLI grade level readability scores (Figure 8-9). These were charted on the same plot area as their measurement unit was the same.

Although the LIX algorithm remains the main tool for analysing this data, the ARI and CLI plots show similar patterns when analysing the same Maltese texts. This analytical process is not designed to investigate the truth of that hypothesis with any certainty and further research and testing would be required to ascertain such a claim. However, the purpose for such an investigation seems to present itself from this data.

Figure 8-9: Text Readability (ARI & CLI) vs. Year



In looking at readability variation, the plots show that for the first five years (1997 – 2001) there was an overall drop in the difficulty level of the Maltese texts based on textual statistics. The LIX and ARI/CLI plots do not follow the exact same pattern over this period, but all three show a discernible drop of similar proportion over the five years. Similarly, all three plots

show that for the following five years (2001 – 2005), the readability levels maintained a similar level, with a slight uptick in the LIX scale of around 3 points. This could be due to slight changes in words per sentence or the number of words with more than 6 characters present in the text.

For the years spanning 2005 – 2010, all three readability formulas show the same pattern of sharp increases and decreases in the readability levels of the Maltese comprehension texts from one year to the next. There are no details in the EAU reports that would explain these sudden variations, however, all three algorithms show similar proportional changes.

In summary, there was a drop in readability from 1997 – 2001, after which the readability of the Maltese texts remained consistent for five years (2001 – 2005). Between 2005 and 2010, the readability levels of the Maltese texts varied between more and less challenging from one year to the next.

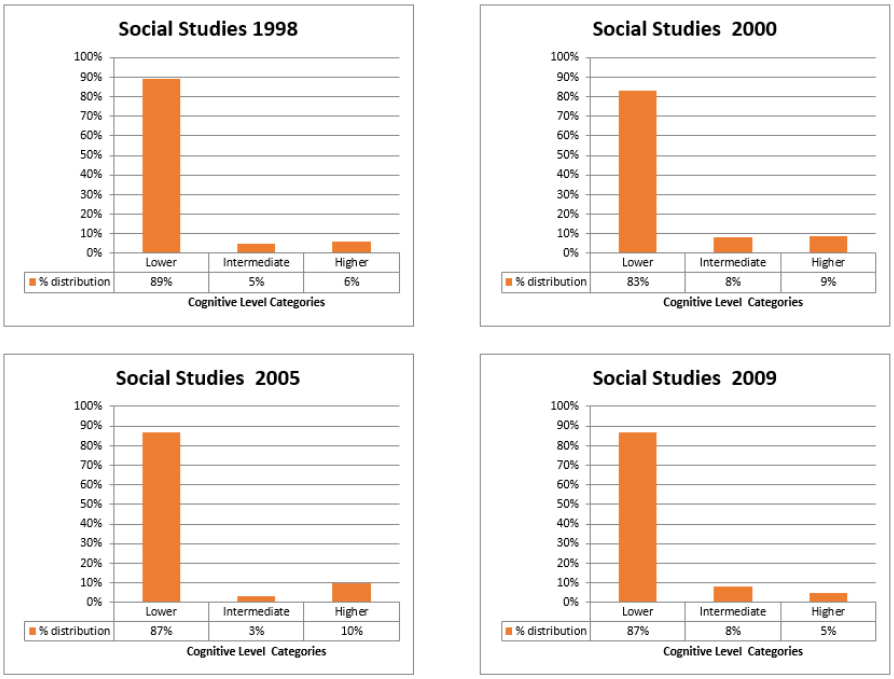
8.3.3 Cognitive item demands — Analysis

The analysis presented in this section is a longitudinal comparison of the percentage weighted score distribution of test items according to question verbs. The comparison covers four specific years (1998, 2000, 2005, 2009) for each of the five subjects. The details of the procedures used to create a cognitive item demand profile are presented in section 6.5.

The cognitive level distribution charts for each of the four years are presented alongside each other and as stated earlier, are not conclusive in determining objective difficulty levels but in establishing an approximation of general longitudinal trends of cognitive level emphasis. Furthermore, the EAU reports included an estimated difficulty level determined by the exam writers for three of the subjects. These are also reviewed at the end of each of the following subsections.

i. Social Studies

Figure 8-10 Question verb analysis: Cognitive demand profiles - Social Studies



The analysis shows that the distribution of cognitive demand of the test forms was centred around lower CL question types working to assess recall and understanding. The CL distribution profile is relatively unchanged between 1998 and 2009, with slight redistribution between intermediate and higher CLs.

The estimated difficulty levels for social studies test forms presented in the EAU reports were established on a three-tier classification system (Easy, Moderate, Difficult) and stated the estimated difficulty distribution of the weighted scores for each form. These estimates are tabulated below in Table 8-13, which shows a continuous distribution to 2005 with a distinct change in the distribution in 2009 indicating a shift to more moderate-level questions.

Table 8-13: EAU report - Estimated difficulty levels - Social Studies

	Easy	Moderate	Difficult
1998	30%	34%	36%
2000	30%	34%	36%
2005	34%	35%	31%
2009	23%	65%	12%

Source: Curriculum Department & Educational Assessment Unit (1998, 2000, 2005, 2009)

A review of the years from 2006 – 2008 (Table 8-14 below) was undertaken to identify when this change took place, and the difficulty levels show that the distribution remained relatively unchanged till 2009. There is no reference in the reports as to what brought about this departure from an otherwise relatively consistent trend.

Table 8-14: EAU report: Estimated difficulty levels - Social Studies (2006 - 2008)

	Easy	Moderate	Difficult
2006	34%	32%	34%
2007	35%	32%	33%
2008	36%	32%	32%

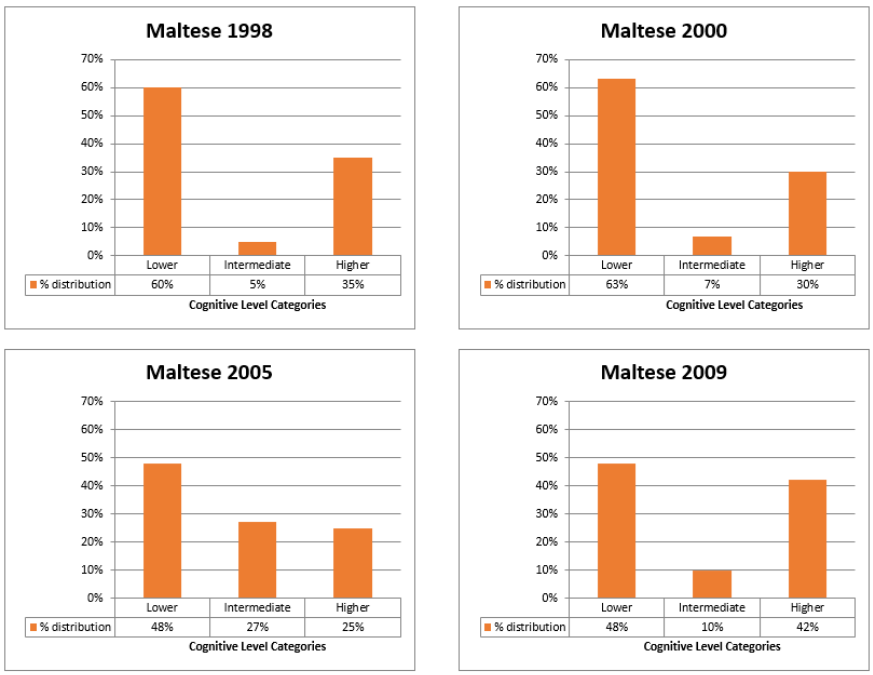
Source: Curriculum Department & Educational Assessment Unit (2006, 2007, 2008)

Nonetheless, the question verb analysis reflects continuity in the cognitive load posed by the four social studies examinations over the period in question (1998 – 2009), suggesting that the cognitive demand remained approximately similar for those test forms over the years.

ii. Maltese

The cognitive demand profile for Maltese retained relative similarity for the 1998 and 2000 test forms, with more emphasis placed on lower cognitive level items. This profile changes in 2005, reflecting greater emphasis on intermediate cognitive demands, and there is another shift in 2009 to incorporate higher CL questions. A review of the other Maltese test forms between 2005 and 2009 identified the introduction of a guided writing exercise in 2008 which required the formulation of a letter resulting in the additional emphasis on higher-level cognitive demand. As essay questions were classified at a higher CL, the introduced essay question led to higher-order writing demands for Maltese test forms starting in 2008. This test item also carried 25% of the mark consistently across the years.

Figure 8-11 Question verb analysis: Cognitive demand profiles - Maltese



Unlike social studies, English, and mathematics, the EAU reports do not incorporate an estimated difficulty level for the Maltese examinations.

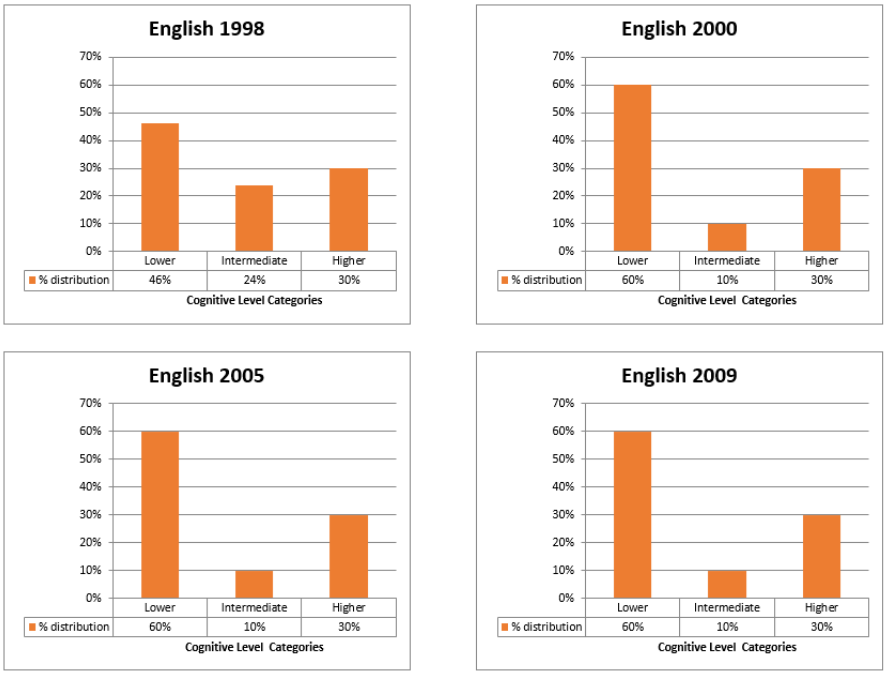
Overall, the cognitive demand posed by the four Maltese test forms increased as the emphasis shifted from lower CL question types to intermediate and higher CL question types.

iii. English

The analysis for English shows that the 1998 test forms put a greater emphasis on the intermediate-level question types than 2000, 2005, and 2009. The 1999 test form demands were similar to 1998 indicating that the change took place in 2000. The 2000, 2005, and 2009 test forms retained a consistent CL profile suggesting consistent application of an examination framework drawn against a fixed test construct.

Like Maltese, the higher CL questions are linked to the essay writing section of the test. In the case of English, this was consistently given a 30% score, and the charts show that there were no other higher CL-type questions added to the mix.

Figure 8-12 Question verb analysis: Cognitive demand profiles - English



The estimated difficulty levels for the English test forms taken from the EAU reports were also structured on a three-tier classification system (Low, Medium, High) similar to the Social Study reports. These stated the estimated difficulty distribution of the weighted scores for each form. The estimates are tabulated in Table 8-15 and, with the exception of 1998 data which was not available, show a shift of approximately 8% from a low to medium difficulty level.

Table 8-15: EAU report - Estimated difficulty levels - English

	Low	Medium	High
1998	NA	NA	NA
2000	33%	40%	27%
2005	29.5%	42.5%	28%
2009	24%	48%	28%

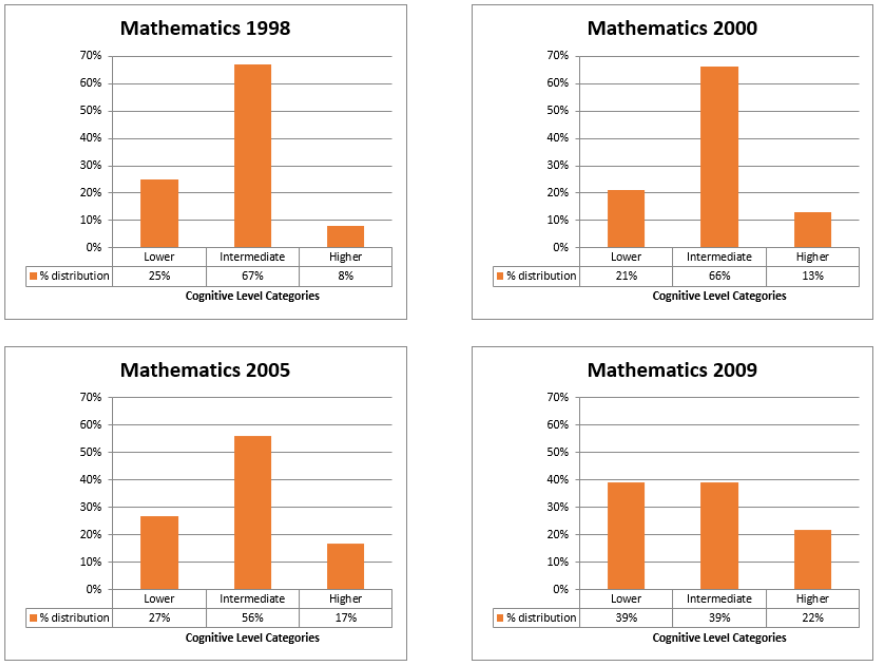
Source: Curriculum Department & Educational Assessment Unit (1998, 2000, 2005, 2009)

The cognitive demand posed by the English test forms from 2000 to 2009 was therefore relatively consistent with no changes to the cognitive demand profile.

iv. Mathematics

Comparing the mathematics test forms for the four years in question shows there were variations across the years with a gradual redistribution of intermediate level question types to lower and higher cognitive levels. Furthermore, compared to the other four subjects, the question types for mathematics have a greater overall distribution of question verbs at intermediate and higher cognitive levels reflecting more application and evaluation type questions than recall.

Figure 8-13 Question verb analysis: Cognitive demand profiles - Mathematics



The 1998 and 2000 test forms show a fair level of consistency. However, there are clear variations between 2000 and 2005, 2005 and 2009. A review of the CL distribution of the question verbs for the years in between indicated that the exam seems to have varied slightly between years but with no discernible trend pattern that would suggest increasing or decreasing difficulty levels.

The estimated difficulty levels for mathematics presented on the EAU reports followed the same three-tier classification system as that used for the English test forms (Low, Medium, High). The estimates are tabulated below in Table 8-16 and show slight variations between medium and high difficulty levels.

Table 8-16: EAU report - Estimated difficulty levels - Mathematics

	Low	Medium	High
1998	27%	41%	32%
2000	28%	46%	26%
2005	27%	42%	31%
2009	27%	45%	28%

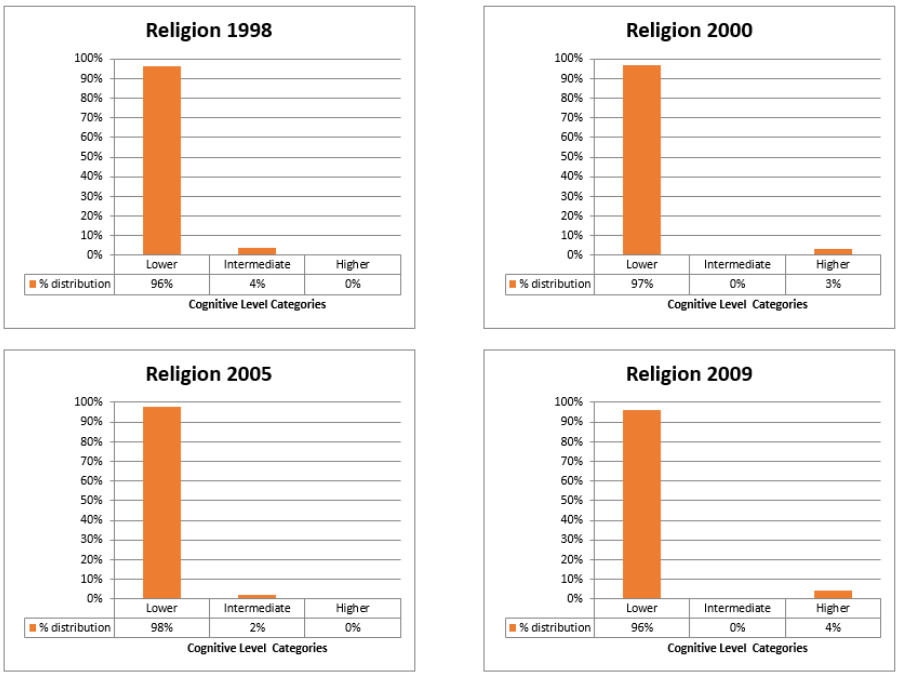
Source: Curriculum Department & Educational Assessment Unit (1998, 2000, 2005, 2009)

As there were no discernible trends, it was not possible to make a longitudinal statement of variation in cognitive demand posed by the mathematics test forms or change in complexity based on this particular analysis.

v. Religion

The distribution of cognitive demand of the religion test forms ran along similar lines to that of social studies being centred around lower CL question types and remaining fairly consistent over the four years. The CL distribution profile is relatively unchanged between 1998 and 2009, with occasional shifts between intermediate and higher CLs.

Figure 8-14 Question verb analysis: Cognitive demand profiles - religion



Similar to the Maltese examination reports, the EAU reports for religion did not incorporate an estimated difficulty for the test forms.

The cognitive demand presented by the religion test forms from 1998 to 2009 was relatively consistent with no major variation in trend for the cognitive demand profile.

Overview: CL profiles analysis

This research recognises that the application of this framework and mechanisms to identify a cognitive level profile of a test form is tentative and cannot be used to draw definitive conclusions about the exact nature of that cognitive profile. It has, however, shown distinct trend differences between the different subjects that reflect the question type characteristics anticipated for those subjects. Social studies and religion tended to draw on recall and understanding type questions, mathematics had a greater proportion of application and evaluation (higher-order thinking) type questions, and the two languages shared similar profile distributions.

In summary, considering the tentative nature of this section of the analysis, what can be drawn from the trend patterns of the CL profiles suggests a consistent difficulty level for social studies, English and religion test forms, an increasing difficulty level for Maltese, and non-determinate shifts in the mathematics test forms.

8.3.4 Format and structures of the test forms and items

This part of the analysis compared the formats and structures of test forms from 1997, 2004, and 2010 (2009 for social studies) to identify possible variations in strategies affecting extraneous cognitive loads of the exams. It was based on a loose framework proposed by Gillmor et al. (2015, p. 6) and structured to establish a longitudinal comparative analysis. The key components of the framework considered translation, visual aids, signalling, weeding, sequencing, aesthetics, and numerical simplicity, discussed earlier in Table 5-1.

The analysis was established as a comparative review of the intermittent test forms, descriptive in nature, and based on longitudinal comparisons of the test item sets assembled by the examiners. Particular attention was given to possible changes in signalling, aesthetics, and paring of questions, items that Gillmor et al. (2015) suggested may be more effective in reducing extraneous factors.

As discussed earlier, for each subject, the analysis processed each of the test forms for the three intermittent years against the framework and provided a general comparative

description of any variations. The framework was not, however, equally applicable to all subjects with sequencing and numerical simplicity relevant only to mathematics (Table 5-2).

Table 8-17: Applicability of strategies to different test forms

Strategy	Applicability to
Translation	All 5 subjects
Visual Aid	All 5 subjects
Signalling	All 5 subjects
Weeding	All 5 subjects
Sequencing	Mathematics
Aesthetics	All 5 subjects
Numerical simplicity	Mathematics

8.3.4.1 Analysis

Analysis of the general format and structures of the test forms sought to first describe the general delivery procedures of the examinations considering if the test forms were delivered in both languages (Maltese and English), if there was a supplementary/resit paper, and if there was a special needs paper. The analysis then proceeded to consider changes in question item strategies as discussed earlier in Table 5-1 and according to Table 8-17.

If a particular change in delivery or strategy was identified in 2004 or 2009/2010, then the test forms from previous years were reviewed to try and identify the point at which the new strategy was introduced.

Social Studies

Delivery procedures: The social studies test papers were offered in both Maltese and English for every session from 1997 to 2009, allowing students to choose a paper according to their preference. The allocated time for each session was consistently kept at 90 minutes. None of the years offered special needs or resits for this examination.

The structures and formats of the two versions remained identical, allowing the analysis of extraneous factors using the proposed framework to be considered for both versions simultaneously by reviewing just one of the language versions of the papers.

Translation: Although the number of pages increased from 9 to 12 between 1997 and 2009, the earlier paper required that the students write an essay about a chosen topic and allowed space for that. This was dropped in 1998 and replaced with fill-in-the-blanks and open-type questions which required more printed pages. After 1998, the word/page count becomes relatively consistent, and the language level also remained consistent as did the question types used.

Visual Aid: Visual aids on these test forms were used as the basis for question items associated with location identification or picture comprehensions. Each exam had 5 or 6 picture-based questions.

Signalling: Signalling to focus attention improved over the years. In 1997, very few cues were used in the paper, and in 2001 the cues became more distinct with the capitalising of particular instructions which were later also written in a bold script or both. Cues on maps varied between numbered points and "black dots" on the map and again to numbered points. Starting in 2008, the cues became more frequent and distinct which reflects an improvement in this particular strategy over time.

Weeding: As the question items remained fairly consistent over the 14 years, the number of extraneous items also remained fairly consistent. No observable changes were made regarding changes to this strategy.

Aesthetics: The aesthetics improved slightly with pictures becoming clearer and a better distribution of question items to allow less cluttering and more white space. The three intermittent years being reviewed had a similar distribution of text and pictures, yet the 1999 paper stood out for having a noticeably higher text density and less white space in the layout.

Overall: The social studies test forms remained relatively similar in their distribution of question items, with slight improvements to signalling and aesthetic strategies. Consequently, none of the changes are considered to have any distinct effect on the extraneous cognitive load exerted by the test forms.

Maltese

Delivery procedures: The allocated time for each exam session was 105 minutes. Special papers were prepared for children of returning migrants or special needs candidates. However, the number of candidates allowed these special papers never totalled more than 9. Furthermore, these special papers were stopped after 2005 and not replaced by an alternative option.

The Maltese test form showed a distinct change between 1997 and 1998 in both structure and format and there was a restructuring in the way grammar was assessed. This change was maintained from 1998 to 2010 increasing the number of questions associated with the comprehension section of the paper. The change also restructured some of the grammatical questions to tie in with the comprehension text and give those items more contextual relevance.

From 2008 to 2010, the linkage between grammar questions and the comprehension text remained but the general structure of the paper changed how the other grammar-related questions were assessed. The new format did away with conjugations and conversions devoid of context and integrated the same assessment constructs into longer paragraphs and sentence structures. This, similar to the previous change in 1998, did away with what Purpura (2013) considers a *"…highly restricted view of the construct."*, and assessed the construct in a broader context-based setting giving a semantic and pragmatic meaning associated with specific situations (2013, p. 4).

Translation: The language level remained relatively similar for all the test forms from 1997 – 2010 with the word/page count remaining relatively consistent.

Visual Aid: The 1997 test form had no visual aids. This changed along with the new strategies employed in 1998, where 7 visual aids were employed to contextualise the question items. Visual aids were included in most of the question items for the test forms leading to 2007. From 2008 to 2010, the visual aids were not used as prolifically as during previous years, dropping from twenty pictures in 2007 to three in 2008, four in 2009, and five in 2010. All visual aids used on the Maltese exams were not associated with spatial constructs but used to render a contextual scenario against which the test subjects were expected to respond.

Signalling: The direct linkage between questions and the comprehension text was supported by clear signalling of what was expected and the location in the text. The other questions also had appropriate signalling strategies applied consistently over the years.

Weeding: The question items remained fairly consistent from 1998 to 2007, and the number of extraneous effecting factors maintained relative consistency with no observable changes being made. From 2008 to 2010, visual aids were reduced, and textual prompts increased to provide more context for the response, however, it was not possible to determine if this increased or decreased the cognitive load overall.

Aesthetics: The overall presentation and format of the test forms remained the same from 1998 to 2010, with text density and white space also remaining approximately similar.

Overall: The Maltese test forms retained a similar distribution of question items as a whole with no significant variations in strategies that would have impacted the overall extraneous cognitive load. The longitudinal changes to the test forms moved toward a contextualisation of the question items through a designed modification to various extraneous factors in 1998 and again in 2008. A detailed understanding of impact on cognitive load cannot be determined from the test forms alone without a controlled investigation.

English

Delivery procedures: The allocated time for each exam session was 105 minutes. Starting in 2009, those students failing the exam were allowed a resit at the end of July. All test forms were structured to begin with language and grammar questions followed by a reading comprehension exercise and a composition.

In comparing the three intermittent test forms (1997, 2004, and 2010), the analysis showed significant changes to parts of the language and grammar sections that required further investigations into when and how those changes had come about. This led to a year-on-year analysis that showed that for all test forms from 1997 to 2010, the first section consisted of ten multiple-choice sub-questions that required underlining or ticking the correct answer. This consistency in the test forms was not, however, reflected in any of the other language and grammar sub-sections which used considerably different strategies to assess similar constructs up till 2005. Starting in 2005, the language assessment structures became more consistent in their assessment strategies.

For the language and grammar question items, the main question type for all years was "*fill in the blanks with the correct word*" type exercises. However, until 2004, the examiners would occasionally include a list of words from which to choose and not do so for the following sitting, requiring the students to derive a suitable word to complete the sentences or phrases.

Translation: The language level for the English test forms remained relatively similar for all the test forms from 1997 – 2010. The number of assessment pages increased from 9 in 1997 to 12 in 2004, then remained the same to 2010.

Visual aid: These were used consistently to add context to most language, comprehension, and composition-type questions. There were no significant variations in the use of this strategy.

Signalling: Each of the sub-sections had the questions and specifications for completion clearly marked at the beginning of the sub-section. Each question was written in bold and highlighted with a border around the question text. An example was always given, and this was done consistently for all test forms.

Weeding: For all the years in question, the text of the questions for the language sections were structured to help students focus specifically on the construct being assessed, except for 2003 and 2004 where an attempt was made to add context to the questions using a storyline.

The 1999 paper had an increase in the text used to assess some of the language constructs. More significantly however, although the language level remained relatively consistent, the

word count for the language section of the test was reduced starting in 2003, thus reducing the extraneous cognitive load.

Aesthetics: The distribution of sections and question items remained the same from 1997 to 2010, and the general aesthetics and layout of the English test forms improved over the years. Starting in 2003, the text density was sharply reduced having less cluttering and more white space. This made the paper less taxing on the reader reducing the extraneous cognitive load.

Overall: The language and grammar assessment constructs varied sporadically and the review of the English test forms over the 14 years showed that the re-distribution of question items and weightings impacted the test forms' ability to support long-term continuity. As argued earlier in the longitudinal construct analysis, this affected the intrinsic cognitive load. These variations also had an inevitable impact on the extraneous cognitive load as the number of questions varied, affecting the average item score as shown in Table 8-6.

The analysis of each sub-section of the English test forms suggests that the variations took place irregularly from year to year — there were intermittent inclusion or exclusion of word lists from which to select the correct answer; the weighted mark for item groups varied; different assessment tools were used to assess similar constructs; and different grammatical constructs were assessed on different test forms. None of these changes were reported or explained in any of the EAU reports.

The analysis of the formats and structures of the English test forms from 1997 to 2010 can only support an argument for partial consistency in the extraneous load from 2005 to 2010.

Mathematics

Delivery procedures: The allocated time for each of the mathematics examination sessions was 105 minutes. Starting in 2009, those students failing the May sitting were allowed a resit at the end of July.

All test forms were structured as a series of questions and sub-questions. The test forms maintained a similar sequence of structures and question formats from 1997 to 2003, which then changed from 2004 to 2006, and once again from 2007 to 2010. The structures and formats became progressively, if only slightly, better organised in terms of layout. The question sequence was also adjusted to retain a similar flow for each of these periods. Even with these adjustments, the flow of the test forms and general question structures remained fairly similar over time. This consistency would have facilitated student preparation and coaching through practising past papers affecting the extraneous load.

Translation: The language level for all three of the intermittent test forms (1997, 2004, and 2010), remained similar throughout. The word count dropped slightly over the years and the question items were distributed over more pages in 2004 and 2010.

Visual aid: The number of diagrams used to support spatial information varied slightly from one year to the next, with the number of spatial diagrams ranging from 11 to 15 and always associated with the same question types.

Signalling: The signalling retained similar strategies of using bold or capitalised text to emphasise cues and focus attention. The 2004 and 2010 had more frequent use of these strategies to identify the values expected to be used in the solution.

Weeding: The different sections became more specific to the construct over time with the questioning strategy having fewer contextual additives and instructions being more direct: "*Work out"; "Multiply"; "Complete"*. In 1997 numeracy was assessed as distinct questions distributed in various parts of the test form but had lost their role as contextual addons by 2004 and were aggregated into a single section. Starting in 2007, the numeracy section had been restructured into a sectional array of ten questions directly assessing the numeracy skills. There were no ambiguities as to what construct was being assessed and as such, the test forms improved slightly.

Sequencing: Each question item had sub-questions and several of these, although associated with the main question, were independent of each other such that the answer to one sub-question was not required to answer any of the following sub-questions. This made each sub-question a distinct assessment item. However, although independent, the sub-question sequence did sometimes become more demanding in terms of process steps needed to solve the problem.

Aesthetics: The 2004 and 2010 test forms had a similar usage of fonts and diagrams. Numbers used throughout the papers used a larger font size compared to the instructional text. The layout, although fairly structured in the 1997 test forms, became better organised in 2004 and even more so in 2010. Each question was sectioned off in the latter test form and clear instructions identified what was expected. The diagrams were also clearer and better defined.

Numerical simplicity: The numbers used over the years remained at the same level. There was some usage of decimals with four significant figures in 1997 as part of an ordering exercise, however, these types of numbers were not used in any of the subsequent test forms from 1998 to 2010. The numbers used seem to have been selected to be simple and straightforward.

Overall: The general format and structures of the mathematics test forms improved from 1997 to 2004 and again in 2010. The changes and improvements were gradual with different strategies being introduced at various times. Numbers were written in larger fonts and diagrams became clearer and slightly more appropriate. The layout of the paper also improved over time with better item distribution making them more distinct. All these were done while maintaining the general test construct consistent for all test forms.

Religion

Delivery procedures: The religion test papers had delivery procedures similar to those of social studies and were offered in Maltese or English for each session from 1997 to 2010. The allocated time for each session was also 90 minutes, and none of the years offered special needs or resits.

The structures and formats of the Maltese and English versions were the same allowing the analysis of extraneous factors to be considered for both versions simultaneously by analysing only one version.

Translation: Although the language level was similar for all three intermittent years (1997, 2004, and 2010), the word count increased, with the number of comprehension texts changing from a single passage in 1997 to two in 2004 and three in 2010. The word count for the whole test form changed from just over nine hundred in 1997 to about one thousand six hundred in 2004 and just over two thousand in 2010. The increase in word counts also stemmed from changes to assessment strategies going from items requiring longer sentence answers in 1997 to more "fill-in-the-blank" items needing single word responses in 2004 and 2010.

Furthermore, although sentence writing was still required in different sections after 1997, each of the sections was subdivided into sets of sub-questions creating a guided response structure. This was not the case in 1997. The change to a more guided strategy seems to have started in 1998 and became more prevalent thereafter. These changes assessed the same constructs but would have reduced the extraneous cognitive load through better structured questions.

Visual aid: These were not commonly used in any of the test forms other than to give a contextual image without affecting the assessment strategy or supporting the response process.

Signalling: Except for underlined questions at the beginning of the sections, signalling was sparsely used in 1997. This improved in 2004 and 2010 with more application of bold text and

examples used to support and guide the respondents. In this respect, the papers improved slightly in 2004 and 2010.

Weeding: The religion test constructs were based on rote learning and recall with a few of the test items requiring a limited form of discussion or critical thinking. Those parts that did require a higher order thinking were fairly broad in scope and although they did not have distractors affecting the extraneous cognitive load, they were not overtly explicit in identifying the construct being assessed. This was common for all three of the intermittent papers reviewed implying the papers remained the same in this regard.

Aesthetics: The aesthetics for the religion test forms did not change much over the years in terms of changing fonts or contextual pictures.

Overall: While the word count and the number of pages increased over the years, the change from unstructured open questions to more guided response structures and fill-in-the-blanks formats had an impact on reducing the extraneous load on the test subjects. The questioning strategies thus improved from 1997 to 1998 to better guide respondents in answering with only minor improvements between 1998 to 2010. The religion examinations can therefore be considered to have improved slightly over the years while remaining true to the original assessment constructs.

General overview of format and structures analysis

The general trend for each of the examination sets was to reduce their extraneous CL by varying the formats and structures of the test forms. These actions improved one or more of the different strategies listed in Table 5-1, and would have reduced the complexity of the test forms, thus impacting the difficulty levels. Social studies, Maltese, mathematics, and religion all showed gradual changes in these strategies to tactically reduce the extraneous load. English on the other hand only seems to have begun implementing such strategies after 2005, with previous years suggesting a less organised approach towards such support mechanisms.

Although this framework worked to determine variations in formats and structures, it cannot however be applied as an independent tool to draw definitive conclusions regarding the impact on the extraneous load. Similar to the previous section, this section is considered to be a supportive component of the overall analysis, informing a broader description of changes affecting the difficulty levels of the test forms.

One more key variation to the implementation structures for these examinations was the introduction of resit exams. Resit examinations were introduced in 2009 for students who had

failed only one examination from the set of five (Curriculum Department & Educational Assessment Unit, 2009, p. 25). This would have had a direct impact on the general examination structures, and subsequently on the overall pass-fail rates for those last two years. However, as will be noted later (9.2.3 Resit Sessions), the statistics of those resit exams were not included in the reported pass-fail rates from the main sitting, and will consequently not affect the analysis of the outputs.

8.3.5 Overview: Cognitive analysis

The analysis presented in the cognitive analysis section was intended to be informative, though not conclusive. The process was applied to understand general longitudinal patterns of change or consistency of the various factors considered to affect the extraneous cognitive load of a test form. The outcomes of this section will later be considered to supplement and support an interpretation of the subject-based outcomes analysis.

Table 8-18 below summarises the cognitive characteristics of each subject and outlines changes that happened over the years.

Table 8-18 Summary table of cognitive analysis

Subject	Readability level	Cognitive Item Demand (CID)	Formats and Structures
Social Studies	N/A	Consistent CID till 2008 Estimated difficulty changed in 2009	No Changes
Maltese	High (1997 – 2000) Moderate (2001 – 2005) Sporadic (2006 – 2010)	Increasing CID (2005 – 2009)	Changes in 1998 and 2008
English	Consistent level except for 1998 & 2001 which had a higher level	Increasing CID (2000 – 2010)	Irregular changes (1998 – 2004) Consistent template (2005 – 2010) Resits introduced 2009
Mathematics	N/A	Redistribution of CID structures starting 2005 from intermediate to Higher and Lower	Continuous improvements (1997 – 2010) Resits introduced 2009
Religion	N/A	Consistent CID throughout	Changes from 1997 – 1998 Slight improvements from 1999 - 2010

8.4 Statistical Analysis — trends in psychometric characteristics

This third and final part of the process and context analysis is designed to understand latent variations in the general examination contexts reflected in psychometric characteristics. It

investigates these variations through a graphical comparison of the item analysis data. The analysis presents longitudinal comparisons by subject but also offers insights into cross-sectional similarities that suggest systematic actions influencing the examination processes.

The analytical methodology applied here was designed in two parts and uses item analysis data drawn from the EAU reports (Curriculum Department & Educational Assessment Unit, 1999 - 2010). The first part reviews the variations in the statistical mean of both facility and discrimination, while the second part establishes a comparative array of D_i vs F scatter plots for each subject and presents a time-series comparison of the plot patterns (Doublesin, 2022). This latter analysis also rendered a more detailed picture of longitudinal and cross-sectional trends.

8.4.1 Part 1: Analysis of statistical mean of D_i and F

This first part of the statistical analysis is a longitudinal comparison of the D_i and F for each subject and is structured around a graphical analysis of the annual statistical mean of these two variables plotted against time in years. As argued earlier in section 3.6, this would reflect on possible variations in the quality and standards of the examinations within the limitations discussed in section 5.3.4.5. Furthermore, section 5.3.4 argued that the combined discrimination index of the complete set of test items can be considered representative of the quality of the examination as a whole.

Average facility and discrimination indices (1999 – 2010) show the statistical mean values for the discrimination and facility indices for each exam from 1999 – 2010. These values were plotted against each year and the resultant plot was reviewed to determine trend changes in examination quality over time. On the graphs that follow — Figure 8-15 to Figure 8-19 below — D_i is represented by the triangular markers while F are the circular markers.

Figure 8-15: Graph of annual statistical mean of D_i and F (Social Studies)

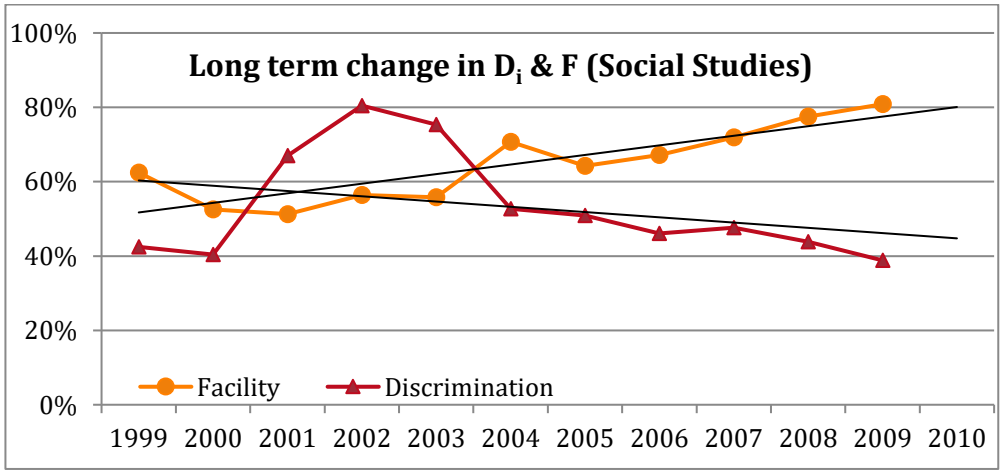


Figure 8-16: Graph of annual statistical mean of D_i and F (Maltese)

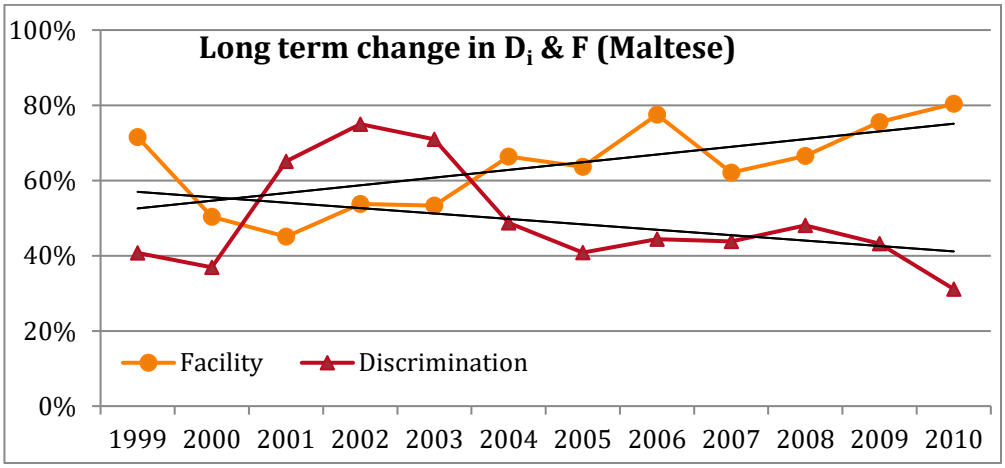


Figure 8-17: Graph of annual statistical mean of D_i and F (English)

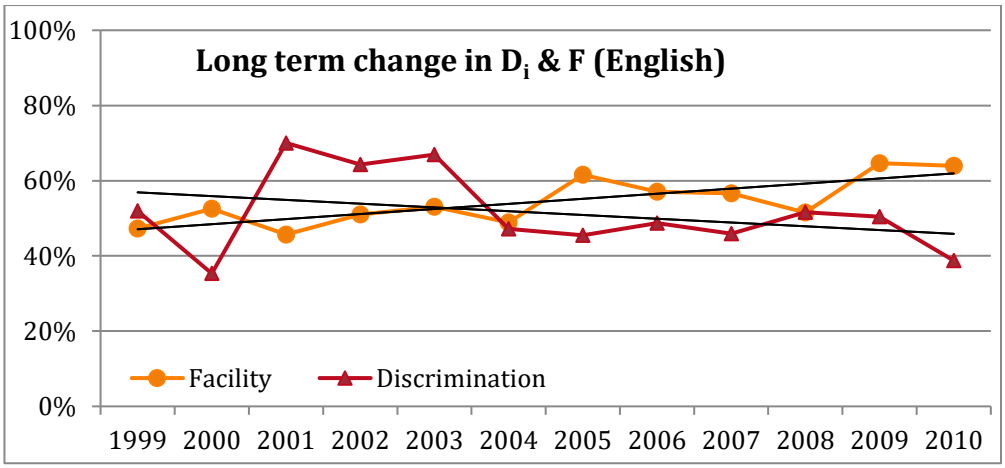


Figure 8-18: Graph of annual statistical mean of D_i and F (Mathematics)

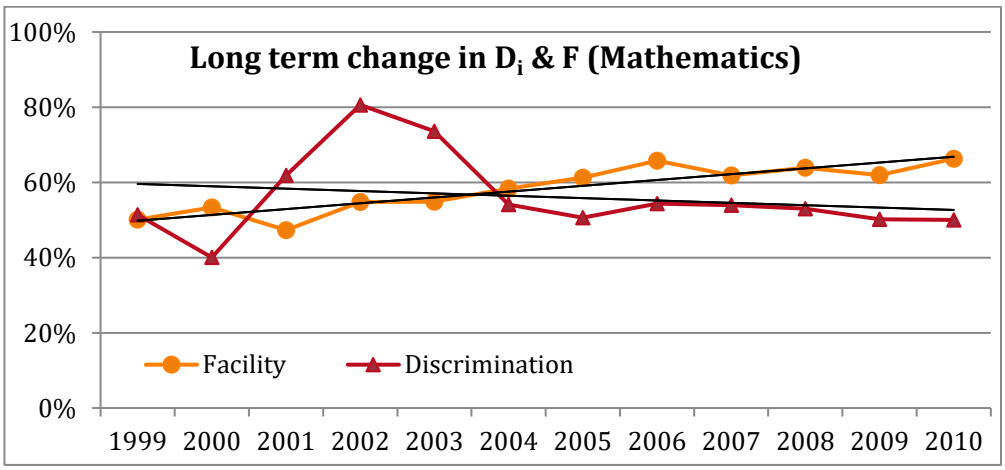
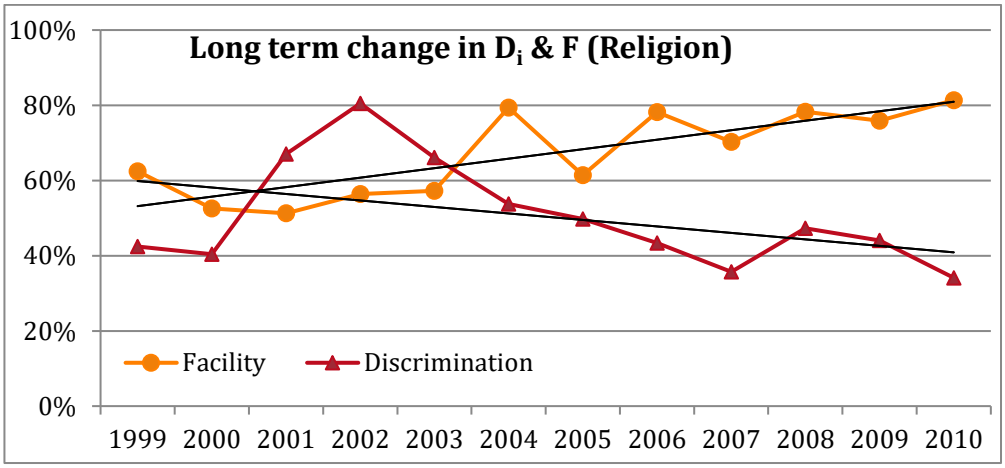


Figure 8-19: Graph of annual statistical mean of Di and F (religion)



Facility

Each of the five sets of graphs was plotted on a common axis for comparative purposes and each had its respective trendline superimposed to support the analysis. As the trendlines represented a generalised rate of change in difficulty level, their gradient gave an approximate measure of that rate of change of test form difficulty. That gradient was estimated using MS Excel's functionality for determining the equation for linear regression and compared across the subjects as shown in Table 8-19. The intercepts, although part of the estimated trendline equation, have no valid meaning in this context.

Table 8-19: F Trendline equations and gradients determined using Excel: (1999 – 2010)

Subject	F Regression equation	Gradient for F
Social Studies	y = 0.0258x - 51.013	2.58% p.a.
Maltese	y = 0.0205x - 40.406	2.05% p.a.
English	y = 0.0135x - 26.519	1.35% p.a.
Mathematics	y = 0.0155x - 30.43	1.55% p.a.
Religion	y = 0.0252x - 49.909	2.52% p.a.

The plots for F show a similar increasing trend from 1999 – 2010, with some idiosyncrasies for each of the subjects.

In looking at the graphical plots themselves:

i. The social studies, Maltese, and religion show a change in pattern starting in 2003, with all three plots showing a more consistent increase for the subsequent 8 years.

ii. English and mathematics have similar plot patterns to each other that differ slightly from the other three, with both having a smaller gradient over the 12 years.

iii. Mathematics has the more consistent rise in F from one year to the next compared to the other subjects, with religion being the more sporadic starting in 2003 and continuing to 2010.

The analysis of the F vs t graphs suggests that all five subject examinations got progressively easier over the twelve years, with social studies and religion seeing the greatest change in facility and English the least. Mathematics also had a decreasing difficulty level similar to the English examinations, however, the set of exams seem to have had more regular year-on-year changes that brought about these changing difficulty levels at a rate of ≈ 1.55% p.a.

Discrimination

A similar analysis was done for the discrimination index and an initial review of the plots showed that except for three years (2001 – 2003), the discrimination power of the examination papers remained fairly constant from 1999 to 2010 for each of the subjects. The overall gradient for the D_i trendlines (1999 – 2010) was estimated and presented below in Table 8-20.

Table 8-20: Di Trendline equations and gradients determined using Excel: (1999 – 2010)

Subject	D_i Regression equation	Gradient for D_i
Social Studies	y = -0.0141x + 28.886	-1.41% p.a.
Maltese	y = -0.0144x + 29.348	-1.44% p.a.
English	y = -0.01x + 20.604	-1.00% p.a.
Mathematics	y = -0.0063x + 13.178	-0.63% p.a.
Religion	y = -0.0172x + 35.058	-1.72% p.a.

The specific actions taken between 2001 and 2003 to increase the discrimination power of the exams have not been identified, however, they clearly had an impact across all five examinations as it shows up as a positive cross-sectional trend. It is likely to have been an institutional level change across the board that affected all exam preparation and is reflected as a common variation in the discrimination power for all exams.

It is apparent from an analysis of the plots that the jump in D_i over those three years affected the trendline gradient, skewing it negatively and suggesting decreasing quality, yet, this would be an incorrect inference. In the absence of those three consecutive years, the rate of change of discrimination power for all five examinations is less than 1% and positive (except for religion), indicating fairly consistent overall quality and implementation of standards for the other 9 years in question.

In summary, the overall statistical analysis suggests that the quality of the test forms remained fairly constant, but the facility as experienced by the students increased (albeit at different rates), meaning that the experienced difficulty level decreased over time.

8.4.2 Part 2: Comparative arrays of Di vs F

In working to understand the psychometric characteristics in further detail, the methodological design considered the application of a comparative array of plots of D_i vs F. As discussed in the methods section 5.3.4, such an analysis was designed to elaborate on how each item on the exam was distributed across the plot area and give a visual understanding of variation in the general psychometric characteristics of each exam. Establishing such a comparative context allowed the longitudinal analysis of year-on-year changes in D_i and F to shed light on instances where outcomes had been affected over shorter periods and trend variations over the longer term. It also presented the possibility of cross-sectional comparative analysis.

The descriptive analyses that follow were structured around the distribution densities of the plots and a comparison of the shape of the trendlines to the ideal-test curve. The arguments made in the methodology section reasoned that the shape of the D_i vs F plots should have retained a relatively similar distribution density and shape for each plot if the standards and quality of the examinations did in fact remain the same (on page 121 above).

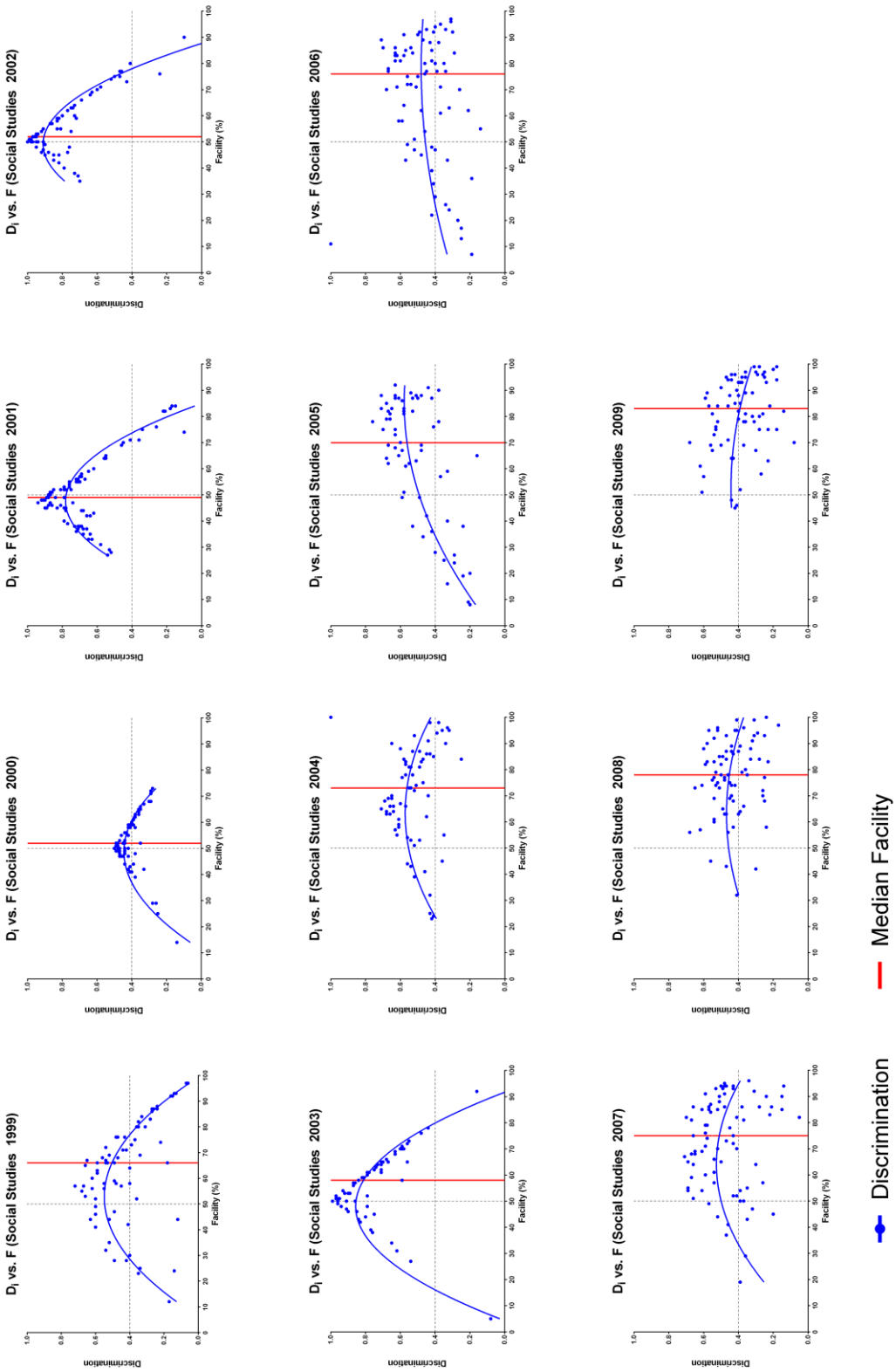
8.4.2.1 Graphical plots of D_i vs F

The D_i vs F plots are presented below by subject and inform the longitudinal analysis that follows. The data used was drawn from the EAU reports (Curriculum Department & Educational Assessment Unit, 1999 - 2010).

As described earlier, the graphical array is presented as five subject rows by twelve yearly columns with each page representing a single row by subject. The plots (Figure 8-20 to Figure 8-24) have a superimposed dotted line at D_i =0.4 and F=50% that creates a quadrant for descriptive purposes, and also includes the median facility to visually highlight any shifting difficulty levels for each subject set. Due to the skewed outcomes of the discrimination index determined in the previous sections, the median for this variable was not included.

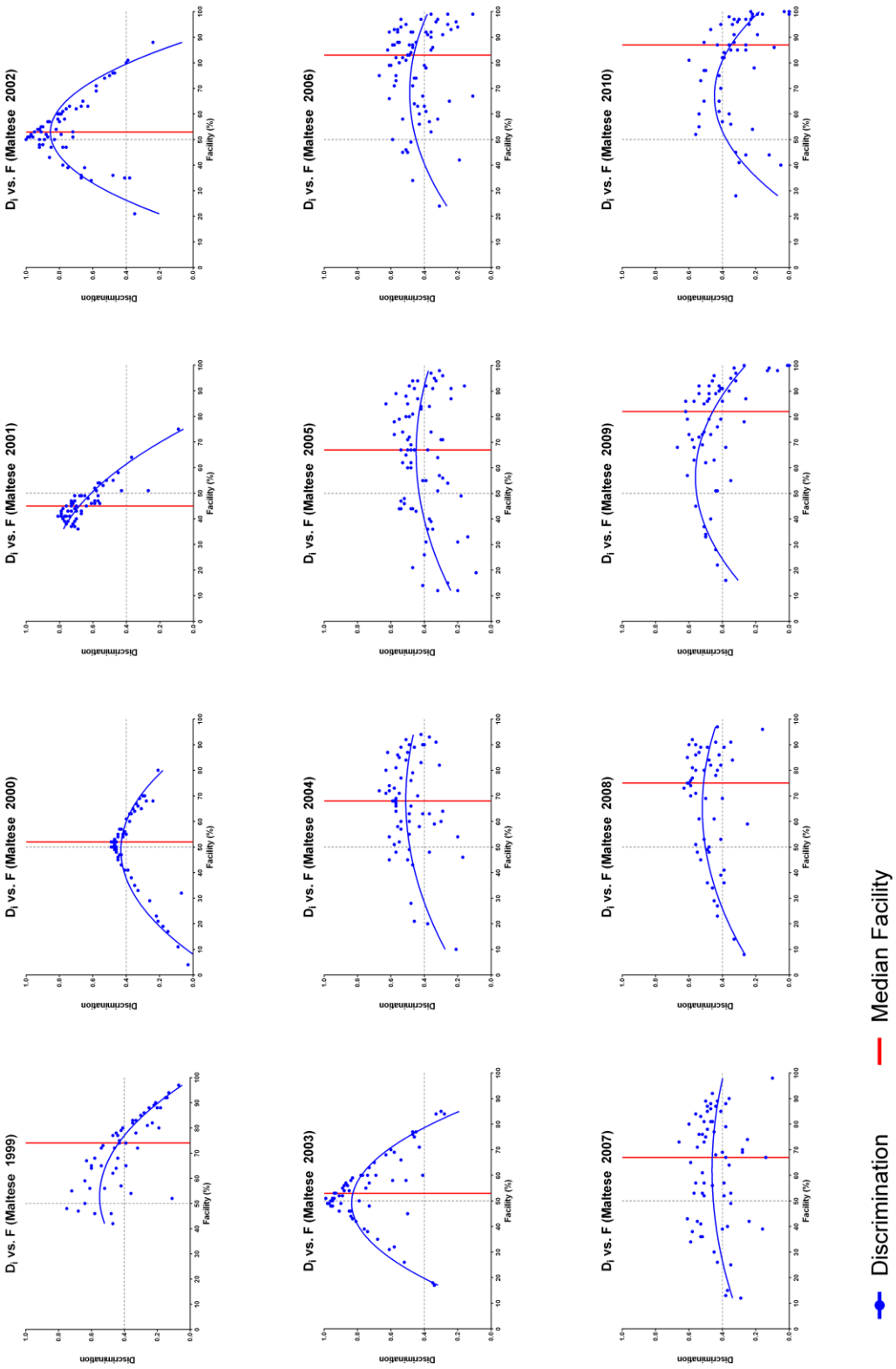
D_i vs F (Social Studies)

Figure 8-20: D_i vs F plots for Social Studies (1999 - 2010)



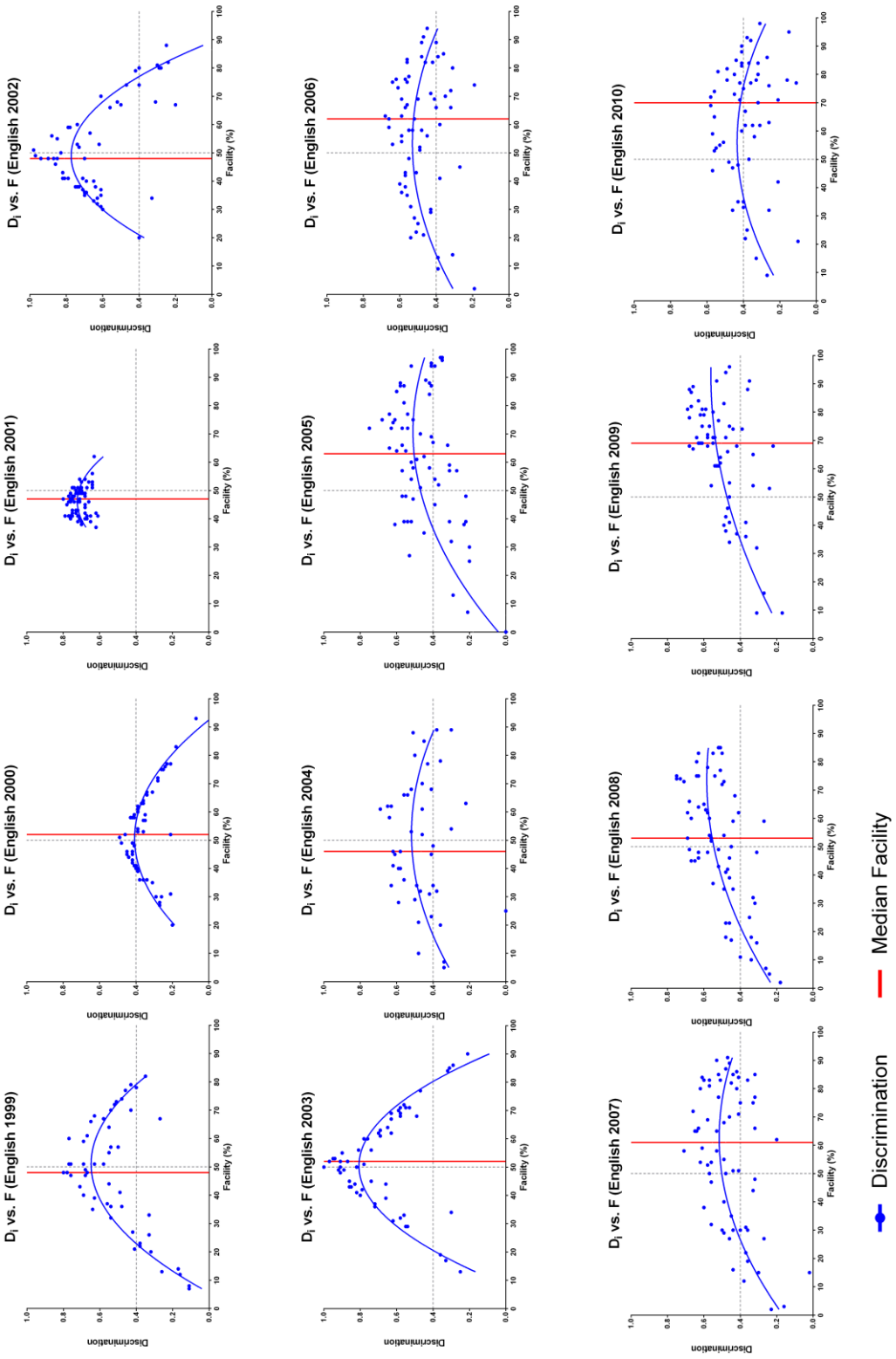
D_i vs F (Maltese)

Figure 8-21: D_i vs F plots for Maltese (1999 - 2010)



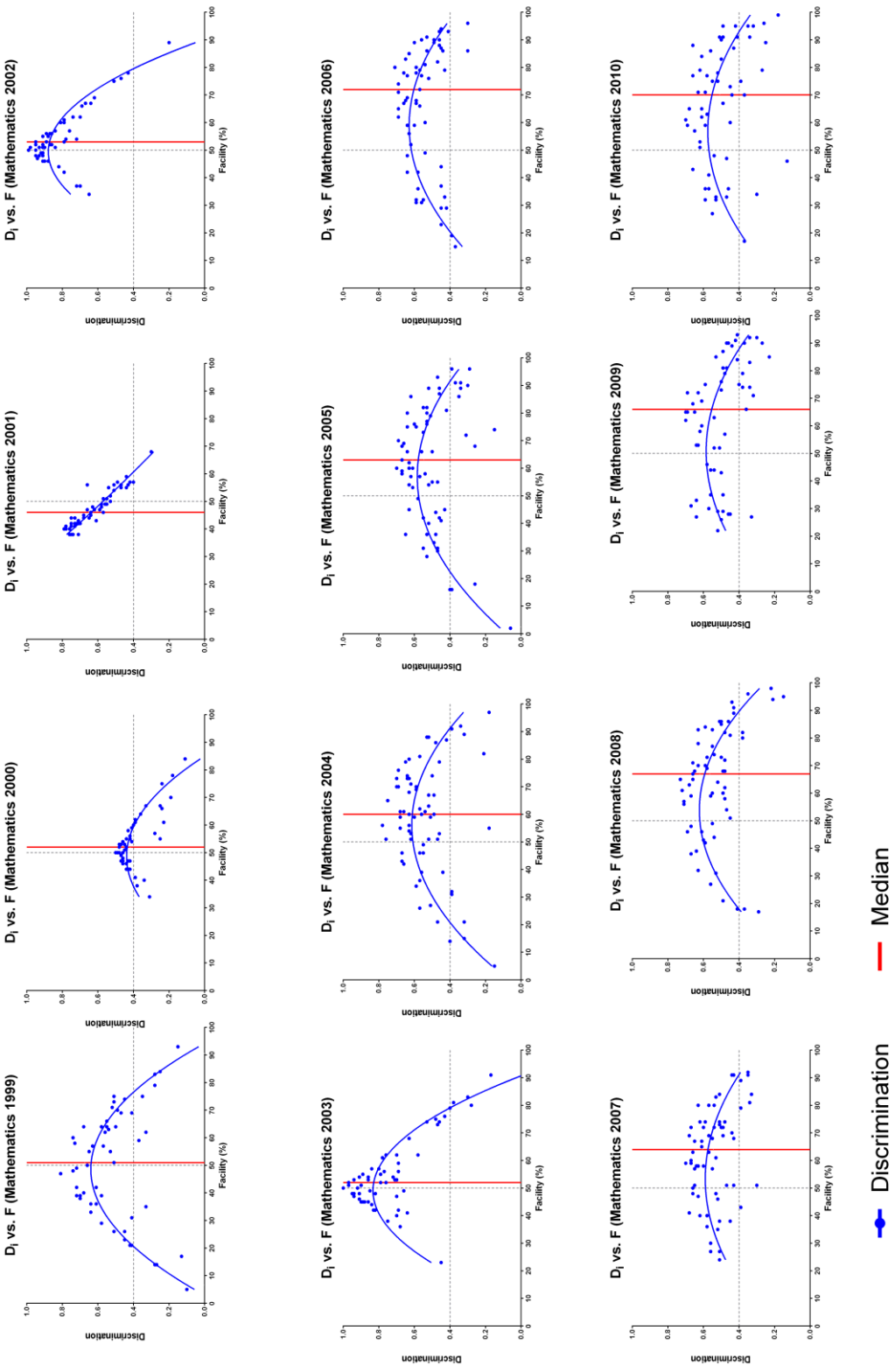
D_i vs F (English)

Figure 8-22: D_i vs F plots for English (1999 - 2010)



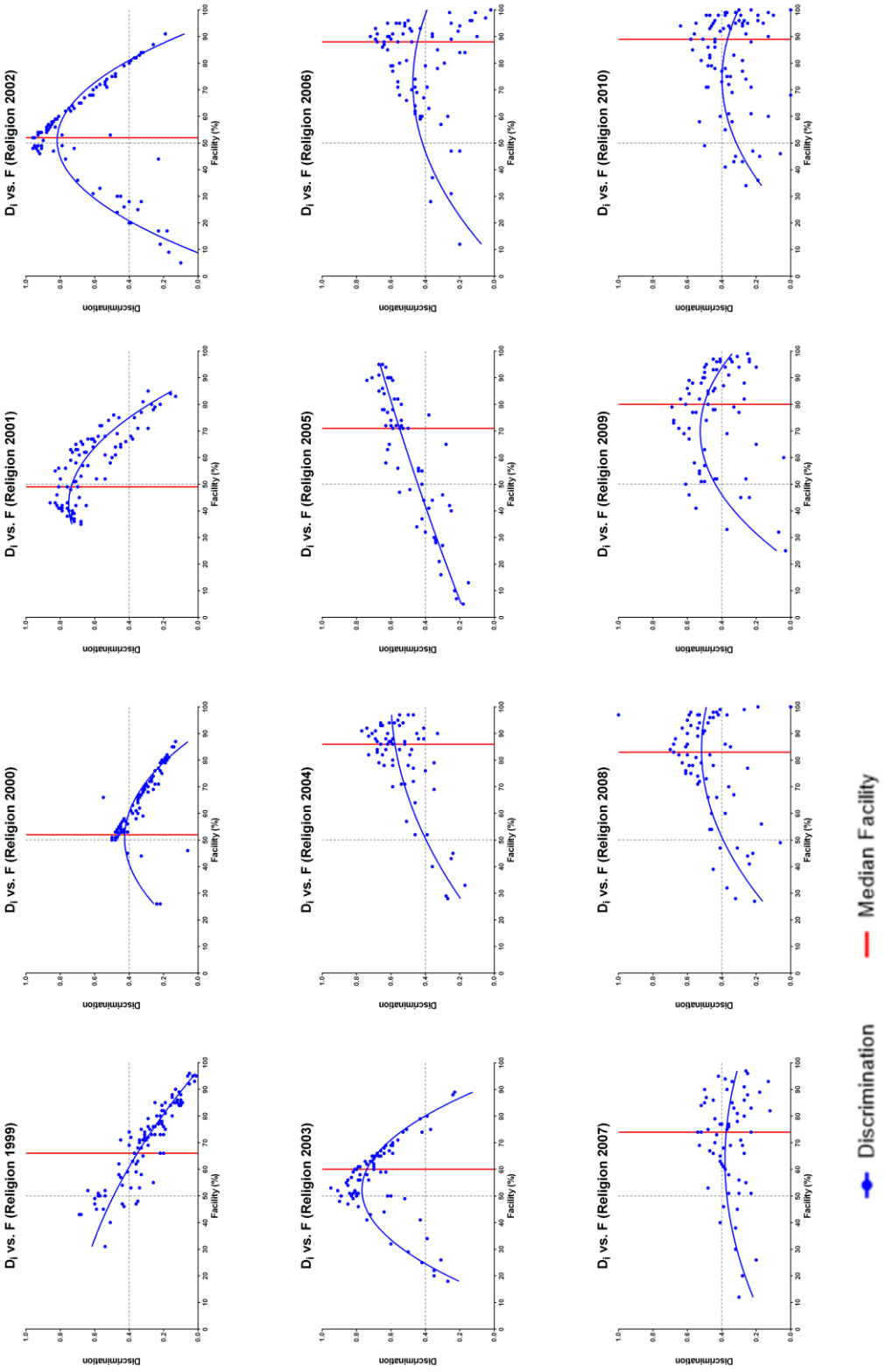
D_i vs F (Mathematics)

Figure 8-23: Di vs F plots for mathematics (1999 - 2010)



D_i vs F (Religion)

Figure 8-24: Di vs F plots for Religion (1999 - 2010)



8.4.2.2 Longitudinal analysis by subject:

Considering each set of plots by subject shows that changes did take place between years and over time. For all five examinations, there is a common and distinct variation in trend between the first 5-year period (1999-2003) and the subsequent seven, which coincides with a change in the administrative director of the curriculum department overseeing the examinations (Curriculum Department & Educational Assessment Unit, 2004).

i. Social Studies

The set of plots representing the social studies examinations (1999 – 2009) show that during the first five years (1999 – 2003), the exam approximated the regular dome-shaped curve expected for an ideal-test situation. Additionally, except for 1999, the distribution density of the plotted points from 2000 to 2003 is fairly tight around the F=50% mark and displays a similar pattern distribution with the points on the right of the apex showing an almost perfect linear regression pattern.

The first five plots in the sequence also reflect consistency in the assessment processes being applied, and although there were minor variations in the discrimination power of the exam from one year to the next, the general test construct seems to have been respected over the five years. Furthermore, from 2001 – 2003 all three examinations retain high overall discrimination.

In 2004, there is a distinct change in general plot patterns. The approximation to the ideal test curve reflected in the first five years is not seen in the last six years as the curves show a distinct flattening. Similarly, the distribution density of the plots takes a sharp shift to the right of the F=50% mark signifying a sudden lowering of the difficulty level. This increased facility is reflected in the jump in median value for each data set, with the plots for the last three years showing a stronger distribution density on the right-hand side (most of the points are squeezed into the two right-side quadrants).

Furthermore, from 2004 onward, the plot density distribution pattern becomes more haphazard compared to the initial five years, and although the statistical analysis of D_i (Figure 8-15: Graph of annual statistical mean of Di and F (Social Studies)) suggests a more consistent value longitudinally, there is a distinct spreading of the discrimination power for each exam.

So, although it can be said that there was continuity in applying the same construct over these six years, it is difficult to conclude that there was any consistency in constructing or implementing the exams.

ii. Maltese

The set of plots for the Maltese language examinations (1999 – 2010) show a similar dual set of patterns with the first five years and the following seven years thereafter reflecting a change in the administration of the exams even though the construct remained the same.

For the first five years, the 2000, 2002, and 2003 sittings approximate the ideal-test curve, with 2000 having a distinctly lower discriminating power than the other two. The trendlines for 1999 and 2001 display what can be interpreted as truncated curve patterns that are uncharacteristic of that 5-year sequence, and indicate a case for deeper investigation as to possible causes. Furthermore, 1999 shows an uncharacteristically high median Facility index with most points plotted to the right of the 50% mark. Another similarity to the social studies set is that during these first five years the points to the right of the apex show an almost perfect linear regression.

In the subsequent seven-year period (2004 – 2010) there is a similar shift in the plot pattern. The curve becomes distinctly flatter and the distribution less uniform. The distribution density of the plots shifts to the right of the F=50% mark and the discrimination power drops again compared to the previous three years. There is a large shift in distribution density to the right of the plot in 2006 before returning to a similar distribution again in 2007, however, the shift is repeated during the last three years (2008 – 2010), reflecting an easier examination overall.

iii. English

Except for 2001, the English examination plots follow a similar trend pattern as the other examinations, approximating an ideal-test curve during the first five years (with low discrimination in 2000) and becoming flatter over the latter years as the distribution density became more sporadically distributed.

The plot for 2001 is atypical of the trend and flags a possible issue for further investigation. The trendline does, on closer inspection, show an approximation to an ideal-test curve albeit on a much tighter x-axis and y-axis range. The trendlines for the years 2004 – 2010 do become flatter, as they did with all other four subjects, however for the English examinations, the distribution density remained constant across the entire horizontal axis and D_i did not vary much over that period.

The examination set associated with the latter seven years can be considered to have retained stronger continuity regarding alignment to the test item development, with the median F generally shifting to the right when compared to the first five years. Except for 2004 and 2008,

the median for each plot (2004 – 2010) stays consistently between the F=60% and F=70% marks. The visual analysis of 2009 and 2010 plots show a slight shift in the density distribution towards the right of the horizontal axis signifying a slight easing in test form difficulty during those years.

iv. Mathematics

The distribution pattern for the set of mathematics examinations is comparable to the English examinations, with similar trends for the first five years followed by a broader distribution of points over the latter seven years.

The plot for 2000 shows a similar drop in D_i as all the other exam plots and 2001 shows an atypical distribution in this 5-year sequence. The trendlines for the years 2004 – 2010 show a similar flattening of the inverted parabolic shape, however, one distinction is that the curves do retain a stronger curvilinear pattern than the other four subject sets. This may be a characteristic associated with mathematics assessments, reflecting a stronger relationship between difficulty level and discrimination power. Escudero et al. (2000) present work that shows that mathematics has a better discrimination power than social science subjects.

Over the last seven years, the examinations for mathematics seem to gradually become relatively easier overall, with the initial five years (1999-2003) showing a more consistent difficulty level closer to F=50%.

Once again, the distinct change in dispersion pattern takes place in 2004 and the set of plots following 2004 reflect a stronger alignment to each other, with similar facility and discrimination power reflecting more consistency in consecutive test forms structures based on the same test construct. As noted earlier, the test construct for mathematics changed starting in 2007 with the introduction of a new syllabus and associated textbooks.

v. Religion

The set of plots for the religion examination is, visually speaking, more pronounced than the other sets in terms of year-on-year changes, with larger variations taking place in between subsequent years. This is also apparent in the statistical analysis (Figure 8-19) above.

The sudden and distinct variation in the trend pattern from 2003 to 2004 is similar to the other subject arrays, changing from a strong approximation of an ideal-test curve to a more irregular distribution.

The trendlines for 1999 – 2003 overall show a strong negative regression to the right of the curve's apex and are similar to the plot sets for social studies. For 1999-2001, although the

computer-generated trendlines are curved, the patterns actually have a strong negative linear regression for the majority of the plotted points. The subsequent two years, 2002 and 2003, retain a predominant negative linear regression for most of the plotted points, however, have a closer approximation to the ideal-test curve as some more challenging questions seem to have been introduced that show a lower discrimination power for a lower F.

The 2004 and 2005 trendlines approximate a positive regression model, but the pattern becomes more curved in subsequent years as the discrimination power of the less challenging question types became less evident. 2005 has a positive linear association and a broad distribution and is distinct from all other plots in this set. The variations in distribution density from 2003 – 2007 reflect broadly changing characteristics in the test forms, and subsequently changing standards for each of those years.

Following 2003, the difficulty level of the religion exam drops drastically compared to the first five years. The average median value for F jumps by approximately 25% points (Appendix A: Average facility and discrimination indices (1999 – 2010)). From 2006 to 2010, the scatter density shifts to the right of the plot, and the discrimination power retains the same distribution across the y-axis — discrimination was not a key factor in these later examinations.

Except for 2007, the patterns for the last five years show a distribution that sees the plot pushing so far to the right that it forms a distinct vertical line along F=1. Overall, the examination shows a major drop in difficulty levels and, similar to social studies, there is a clear spreading of the discrimination power for each exam suggesting a change in the associated standards.

8.4.2.3 Cross-sectional comparisons and combined longitudinal comparisons

The common structure of the five array sets allowed for cross-sectional comparisons between subjects for each year. It also allowed for a combined longitudinal comparison of the five-yearly plots to identify common trait variations that would reflect on more systemic influences affecting the whole examination process. This section combines the cross-sectional and longitudinal cross-sectional analysis due to common descriptive threads.

As noted in the previous section, there are different degrees of variation for each subject on a year-to-year basis. However, there are striking cross-sectional similarities across the five subjects for each of those years. Several of these commonalities observed in the aggregated sets can be identified as particular and are listed here.

- The trendlines of the graphs for the first five years approximate the ideal-test curve in most cases, with 2000 showing a distinct drop in discrimination and 2001 being atypical to the sequence. From 1999 to 2003, the patterns show a distinct dome-shaped trendline or varying segments of a dome shape that approximates the ideal-test curve. In the subsequent seven-year period, the trendlines become distinctly flatter overall. As stated earlier, this coincides with a change in the administrative director of the curriculum department and is likely to reflect a change in assessment policy and procedures.
- In 2000, all five exam plots show a general decrease in discrimination power which increases again during the subsequent three-year period. There is no indication in the reports of what caused this drop in D_i for 2000.
- Except for religion, the plots for 2001 show a relatively dense aggregation of points around the F=50% mark that is atypical to this 5-year set. The EAU report does not give any indication of modulation of the results for these exams, however, the similarity in distribution indicates that this could be an explanation.
- The patterns for the last 5 years (2006 – 2010) show a shifting distribution density towards the right of the Facility axis leaving more distinct "white space" on the left side of each graph. This pattern is more distinctive for the social studies and religion tests than English, mathematics, and Maltese, and reflects a decreasing difficulty level for those exams.
- From 2004 to 2010, English, Maltese, and mathematics display a broader distribution of points than the other two examinations. However, as noted earlier, although they have a similar distribution to each other, the mathematics examinations show a stronger curvilinear relationship between D_i and F which may be a property associated with the nature of mathematics assessments compared to language assessments. The similarity in patterns for these three examinations indicates that the assessment standards were more consistent over the years than for the other two exams.
- The distribution density for both religion and social studies shows a marked displacement towards the right side of the horizontal axis between 2004 and 2010. This reflects an easing of the difficulty level of these two examinations associated with a possible change in assessment policy. The construct for each of these exams remained the same throughout the 12 years, however there were changes in the assessment tools that reduced the capacity to measure higher cognitive abilities, making the assessment more lenient in its purpose.

- Looking at the distribution densities relative to the superimposed quadrants (D_i =0.4, F=50%) for the last seven years of each subject set shows an increasing shift in dispersion towards the right-hand side quadrants compared to the first five years. The distribution densities are thinly spread across the D_i = 0.4 mark and mostly located in the space F>50%. Very few questions had a high difficulty level, and this shift reflects the changes in test form characteristics over the years. Consequently, this raises questions about the suitability for purpose of the examination as a benchmark exam.

8.4.3 Overview: Statistical analysis

The statistical analysis was structured as part of the broader investigation into the process and context associated with the JLEE and the longitudinal consistency and continuity of the test forms. The two elements of this section of the analysis suggest that the quality and standards of the different subject exams changed intermittently and by varying degrees. Social studies and religion appear to have become easier over time, with Maltese showing moderate easing in difficulty levels, and English and mathematics showing the least change.

While these changes can be quantified through a determination of the mean for D_i and F, they were not a determining factor on their own. Rather, and as part of the contextual analysis, these average values together with the D_i vs F plots rendered a picture of shifting psychometric characteristics that implied changing standards. At this point, it would not be pertinent to make any definitive statements about the associated quality of the exams without integrating information from the cognitive analysis. This will follow in subsequent sections.

One further observation was that although the statistical analysis shown in Part 1 (Figure 8-15 - Figure 8-19) suggests a longitudinally consistent discrimination power for all subjects, the subsequent D_i vs F array in part 2 (Figure 8-20 - Figure 8-24) show that the distribution of D_i became more spread out on either side of the trendline. The implication is that the discrimination power of test items became more sporadic as those items become less difficult.

8.5 Summary: Process and context analysis

The process and context analysis were structured to better understand variations in complexity of the test forms and reflect on any changes in their quality, standards, and difficulty levels. Its framework was developed to support the first research question so that its outcomes could inform the second. This relied on determining trend variations in intrinsic and extraneous cognitive loads of the test forms, and a statistical analysis of the psychometric properties of the exams.

The multitude of factors influencing the complexity of test forms therefore required a range of analytical mechanisms (Figure 8-1) to consider different affecting aspects. These were organised into cognitive and statistical analytical processes, with the combination of their longitudinal comparatives subsequently informing the overall contextual analysis. The information drawn from these two analytical methods has led this research to make use of the statistical analysis to infer possible variations in complexity and difficulty levels of the exams, and variations in discrimination power associated with quality. The cognitive analysis, on the other hand, was used to further inform any interpretations associated with sudden or irregular variations in the longitudinal analysis of outcomes discussed in the following chapter. While the statistical analysis of the psychometric characteristics could offer more definitive insights, the cognitive analysis was not as clear-cut in determining overall variations in quality, but did highlight specific change events that reflected changing standards. Table 8-21 is a summary of these changes.

Furthermore, the two analytical systems did not necessarily complement each other in their findings. Looking at social studies, the cognitive analysis indicates that there was consistency and continuity throughout the thirteen years that it was administered. There was no specific change to indicate any variations in the construct or test form complexity and difficulty levels. However, the statistical analysis shows that the difficulty levels dropped at an average rate of 2.58% p.a. over a twelve-year period. A similar rate of change was indicated for the religion examinations which also had little to suggest reasons for the change from the cognitive analysis. The cognitive analysis for Maltese on the other hand showed an increase in cognitive demand, but the statistical analysis showed a moderate easing in experienced difficulty levels.

In summarising the analysis of process and context, the general understandings drawn from both analytical procedures indicate that there were observable variations in the contextual landscape over the fourteen years. The general trends, summarised below in Table 8-21, indicated that:

i. The validity of the exam remained constant and continuous for social studies, Maltese, and English, and changed in 2007 for mathematics. Continuity for religion was indeterminate.
ii. Construct consistency varied slightly for social studies and mathematics; saw a two-year variation in marking distribution for English (2002 & 2004) before returning to a previous model; and was indeterminate for Maltese and religion.

iii. All the exams became relatively easier as reflected by the pattern shifts in the D_i vs F arrays for each subject. Social studies and religion underwent the greatest variation in complexity while English and mathematics the least.

iv. The quality of all exams reflected by the discrimination power remained relatively consistent, except for three years (2001 – 2003). However, the later years showed a broader variation in item quality albeit retaining a similar average discriminating power.

v. The general psychometric changes over the 12 years (1999 – 2010) followed similar cross-sectional fluctuations.

vi. Different subjects had specific changes at separate times related specifically to that subject's test constructs or forms that were not necessarily reflected in the statistical analysis.

Table 8-21 Cognitive analysis summary

	Linking Constructs			Formats and structures	
Subject	Continuity (1997 – 2010)	Consistency	Readability	Cognitive Item Demand (CID)	Formats and Structures
Social Studies	Yes: Continuity Maintained	Yes: Minor Fluctuations (inconsequential)	N/A	Consistent throughout / Estimated difficulty changed (2009)	No Changes
Maltese	Yes: Continuity Maintained	No Information available	High (1997 – 2000) Moderate (2001 – 2005) Sporadic (2006 – 2010)	Increasing CID (2005 – 2009)	Changes in 1998 and 2008
English	Yes: Continuity Maintained	No: Construct changes in 2003 and 2004 only (score redistribution). Consistent across all other years	Consistent level except for 1998 & 2001 which had a higher level	Increasing CID (2000 – 2010)	Irregular CIG (1998 – 2004) Consistent CIG (2005 – 2010) Resits introduced 2009
Mathematics	Yes: Continuity Maintained (1997 – 2006) Syllabus change (2007 – 2010)	Yes	N/A	Redistribution of CID structures starting 2005 from intermediate to Higher and Lower	Continuous improvements (1997 – 2010) Resits introduced 2009
Religion	Not conclusive	No Information available	N/A	Consistent throughout	Changes from 1997 – 1998 Slight improvements from 1999 - 2010

9 Outputs: Outcomes and achievement analysis

9.1 Chapter Overview

This chapter presents a structured analysis of student outcomes and aims to respond to the third research question in determining variations in those outcomes before establishing if they could be associated with the policy introductions. It will also work to understand if there was any impact that would reflect a change in the quality of education across the system.

The first section reviews the student populations in terms of numbers (rather than demographics) from 1997 – 2010 and considers variations in those populations as well as minor discrepancies between the reports and the records.

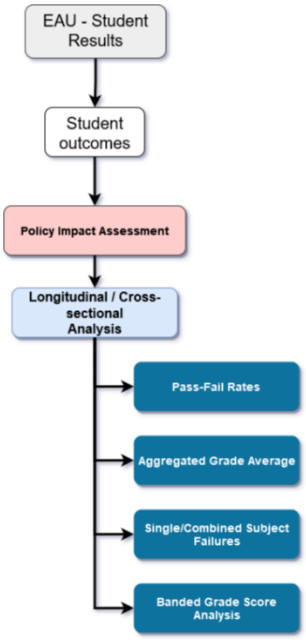
The second part of this chapter looks directly at variations in the pass-fail rates over that same period and compares any changes in trends to information drawn from the analysis of policy, the analysis of contexts, and cognitive variations in the examination papers to determine possible influencing factors.

The third section considers the combined scores and presents an analysis of a compensatory set of outcomes (Douglas & Mislevy, 2010; McBee et al. 2014) by aggregating the five scores together in order to reduce any skewing influences resulting from the EAU's conjunctive criteria.

The final two sections analyse subject-specific outcomes. The fourth part looks at banded scores (A+B+C and D+E) to understand if there were any longitudinal or cross-sectional variations in grade proportion distribution for each subject. The last section then considers each set of subject examinations independently to determine the influence of their particular outcomes on the general rates of success over the years.

The different analytical steps that make up the outcomes analysis part of the research are outlined in Figure 9-1 below and form part of the broader analysis flow diagram (Figure 4-1). Furthermore, part of the analysis required an extended investigation into trend variations dating back beyond the selected period 1997 – 2010 and used data from the reports to extend the timeline back to 1988. This extension was included to add to the historical context of change and showed that improvement spurts in achievement became more common between 2000 and 2010.

Figure 9-1 Outcomes analysis flow diagram



9.2 Population and cohort numbers

Population data was drawn from the examination reports and the record of results discussed earlier in section 4.4.2. The first was the official EAU reports, which included a statistical detail of subsequent Year 6 populations and applicant cohorts and are referred to as “the reports”. The second was the aggregation of result records in terms of subject-based grade scores and Pass-Fail outcomes which are referred to as “the records”. Both reported and recorded populations and cohort statistics are presented in Table 9-1 Year VI student population and cohort numbers by year and compared.

The second column in Table 9-1 shows the total number of Year 6 students in all state and non-state schools across Malta and Gozo, and represents the entire population eligible to apply for the JLEE. Column 3 represents the reported number of applicants by year (Curriculum Department & Educational Assessment Unit, 2010, p. 10). The reported applicants include state, non-state, and secondary school students who applied to sit for the JLEE. The fourth

column presents the count of results taken from the outcomes record by counting the total number of pass or fail results. The final two columns represent the difference between reported and recorded numbers (showing some discrepancies), and the percent of eligible individuals actually sitting for the exams during that year.

9.2.1 Discrepancies in the reported and recorded data

The Pass-Fail count (column 4) should have had the same value as the applicant count (column 3), but there were some discrepancies in the record of results and the reported applicant numbers. This was not due to attrition as those scores were included in the record as absentees. There are a couple of possible explanations for the difference as all the differences shown are negative signifying fewer outcomes records. Firstly, for the years where the difference is greater than 0.1% one or more record pages may have been missed during the scanning of the records. The second explanation for those years that had a difference of less than 0.1% is that one or more record lines may have been cropped during the digitisation process, thus losing the data for that record.

Table 9-1 Year VI student population and cohort numbers by year

Year	Population	Reported applicants	Record count	Reported – recorded (% difference)	Record count / population (%)
1997	6174	4592	4592	0.00%	74%
1998	N/A	4602	4574	-0.61%	N/A
1999	6237	4656	4656	0.00%	75%
2000	6319	4732	4732	0.00%	75%
2001	6255	4600	4600	0.00%	74%
2002	6217	4547	4514	-0.73%	73%
2003	6078	4503	4503	0.00%	74%
2004	6070	4394	4350	-1.01%	72%
2005	5898	4302	4294	-0.19%	73%
2006	5583	3968	3965	-0.08%	71%
2007	5190	3772	3771	-0.03%	73%
2008	5116	3625	3624	-0.03%	71%
2009	4970	3197	3187	-0.31%	64%
2010	4412	2990	2988	-0.07%	68%

Source: Curriculum Department & Educational Assessment Unit, 1997 - 2010

The record data for 1998 was missing the eligible student population numbers, however EAU reports show that those numbers remained approximately the same as the previous and subsequent years (Curriculum Department & Educational Assessment Unit, 2000, p. 7).

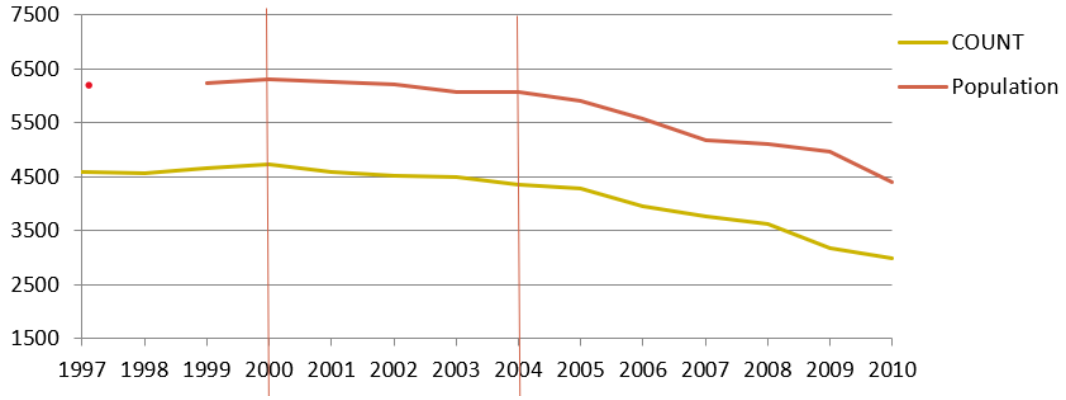
Maintaining a digital record, as opposed to hard copies, would work to minimise these discrepancies by minimising errors made during the digitisation process. Attempts to identify

missing records for those differences >0.1% proved futile without direct access to the hard copies themselves.

9.2.2 Decreasing student population

The number of students actually sitting for the JLEE over the fourteen years decreased gradually over the years. Figure 9-2 shows that this drop matched the change in the eligible student population across the islands and reflects a declining student population in Malta and Gozo at the time.

Figure 9-2 Student population and JLEE record numbers over time



Yet, the percentage of eligible individuals applying and sitting for the JLEE remained fairly consistent till 2008, averaging 73% over this period.

9.2.3 Resit Sessions

The Ministry introduced a resit session in 2009 for those students who failed a single exam during the first sitting (Curriculum Department & Educational Assessment Unit, 2009, p. 25). Considering that all exams needed to be passed with a minimum grade of C, the outcomes of this resit would have had an impact on the overall pass rates (Kickert et al. 2021). However, the number of candidates who were successful the second time round was not included in the record of results that are being used in this study. An analysis of the record of results data matched the reported statistics for the first session in the respective EAU reports.

A brief review of the impact of introducing the resits will be discussed in this subsection with the tables below reproducing the statistics reported in the EAU reports.

Table 9-2 Pass rates for the first and resit sessions (2009)

Pass Rate for the First Session	67.1% (2143 out of 3197)
Pass Rate for the Resit Session	57.7% (213 out of 369)
Combined Pass Rate	73.7% (2356 out of 3197)

Source: (Curriculum Department & Educational Assessment Unit, 2009, p. 26)

Table 9-3 Pass rates for the first and resit sessions (2010)

Pass Rate for the First Session	67.1% (2008 out of 2990)
Pass Rate for the Resit Session	34.9% (114 out of 327)
Combined Pass Rate	71.0% (2122 out of 2990)

Source: (Curriculum Department & Educational Assessment Unit, 2010, p. 23)

The increased pass rate of 6.6% in 2009 and 3.9% in 2010 had a substantial impact on overall outcomes and would have shown up on the impact assessment if included in the record of results. As the successful resit candidates were not however included in that record, nor were their updated grades for the failed subject, the analysis remained true to the original baseline of analysing data from the first JLEE session.

9.3 Pass-Fail rates

This section analyses the longitudinal changes in achievement statistics, cross-referencing them with the analysis of changes in policy, contexts, and cognitive loads by overlapping the timelines from the different analytical processes. The analysis of outcomes as a pass-fail percentage ratio played a key role in determining the impact of policy introduction. It relied on percentage pass rates, percentage fail rates, and the difference between the two being plotted against time.

Table 9-4 shows the overall pass-fail rates of JLEE applicants for the period being considered for this study (1997 – 2010) based on the ministry's criteria for success. This is followed by a second table showing the pass-fail rates for the period from 1988 – 1996. The extended period based on reported data determined trends prior to the principal period of analysis to understand if there were similarity between trends before and after 2000 and informing a broader longitudinal analysis. The plotted data is shown in Figure 9-4 Pass-Fail rates (1988 - 2010).

Table 9-4 Overall Pass/Fail rates (1997-2010) based on Ministry criteria

Year	Record Count	Passed	Failed	% Passed	% Failed	Δ %
1997	4592	2393	2199	52.1%	47.9%	4.2%
1998	4574	2392	2182	52.3%	47.7%	4.6%
1999	4656	2353	2303	50.5%	49.5%	1.1%
2000	4732	2447	2285	51.7%	48.3%	3.4%
2001	4600	2516	2084	54.7%	45.3%	9.4%
2002	4514	2437	2077	54.0%	46.0%	8.0%
2003	4503	2487	2016	55.2%	44.8%	10.5%
2004	4350	2365	1985	54.4%	45.6%	8.7%
2005	4294	2572	1722	59.9%	40.1%	19.8%
2006	3965	2387	1578	60.2%	39.8%	20.4%
2007	3771	2300	1471	61.0%	39.0%	22.0%
2008	3624	2202	1422	60.8%	39.2%	21.5%
2009	3187	2137	1050	67.1%	32.9%	34.1%
2010	2988	2008	980	67.2%	32.8%	34.4%

Source: (EAU : Educational Assessment Unit, n.d.)

It is noted here that the analysis of context associated with the extended period was not part of this study, however, their consideration remains relevant to inform the context as the purpose of the JLEE remained the same for both periods before and after 1997. Grima et al. (2008) have pointed out that various measures have been taken over the years to improve the examination system (2008, p. 100) and the extension will reflect possible impacts that took place before 1997.

Table 9-5 Overall Pass/Fail rates (1988-1996)

Year	Record Count	Passed	Failed	% Passed	% Failed	Δ %
1988	4188	1960	2228	47%	53%	-6%
1989	4096	1944	2152	47%	53%	-5%
1990	4342	1892	2450	44%	56%	-13%
1991	4273	1962	2311	46%	54%	-8%
1992	4425	1988	2437	45%	55%	-10%
1993	4769	2153	2616	45%	55%	-10%
1994	4812	2440	2372	51%	49%	1%
1995	4743	2346	2397	49%	51%	-1%
1996	4558	2409	2149	53%	47%	6%

Source: (Curriculum Department & Educational Assessment Unit, 2010)

9.3.1 Graphical presentation of pass-fail rates (1988 – 2010)

Figure 9-4 Pass-Fail rates (1988 - 2010) shows seven separate plots divided across the introduction of the NMC in 2000 with each plot represented by a colour and symbol as described in Figure 9-3 Pass-Fail rates legend.

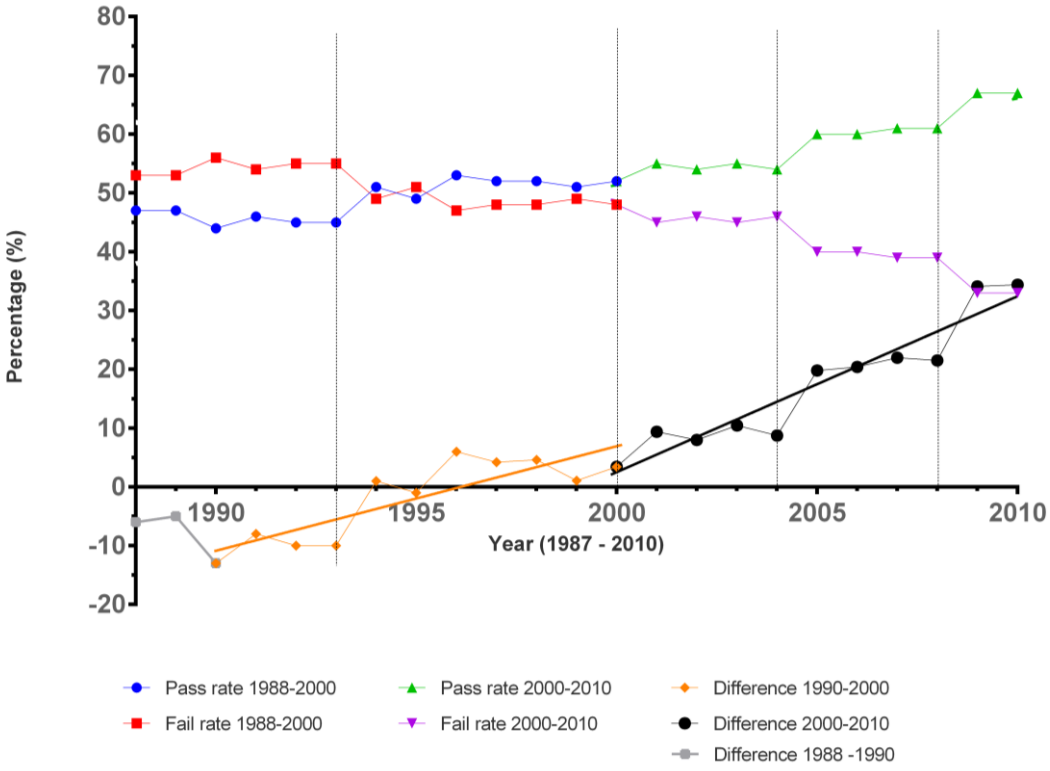
Figure 9-3 Pass-Fail rates legend

The aggregation of the plots facilitated the longitudinal comparisons that underpin this analysis and shows a continuous change in outcomes over the 23 years with intermittent jumps in the trend. Four vertical lines have been overlayed on the plot to show where atypical changes in trend had taken place. These correspond with 1993, 2000, 2004, and 2008.

The difference in pass-fail rates from 1988 to 1990 is plotted separately as it was not included in the gradient calculation for $\Delta\%$. Although 1988 did see the introduction of social studies and religion in the examination set (Grima et al. 2008), and as such made them comparatively relevant, this intended comparison only considered the decade before and after the introduction of the NMC to present a consistent time sequence in determining rates of change.

The plots of difference between the pass and failing percentages (1990 – 2000) and (2000 – 2010) include the trend lines for $\Delta\%$ for which the gradients were determined.

Figure 9-4 Pass-Fail rates (1988 - 2010)



9.3.2 Analysis of pass-fail rates

This section presents the analysis of the graphical representation shown in Figure 9-4 and describes the general trends across the two decades.

The pass and fail rates from 1988 to 2000 show a gradual improvement in outcomes with an observable increase between 1993 and 1994 of 6%. The plots then retain relative consistency until 2000. The jump shown from 1995 to 1996 is not sustained in the long term as the following data displays regression to the previous 1994 levels. There is once again a noticeable jump of 3% in success rates from 2000 to 2001, 5.5% between 2004 to 2005 and 6.3% from 2008 to 2009, giving a staggered look to the improvements. The plots do not show any signs of backsliding back towards the mean for any of these upticks in pass rates suggesting a sustained level of improvement in outcomes.

The segmented regression seen on this plot is what Collins (2006) describes as *"discontinuity in continuous change"* or *"piecewise growth model"* that would be typical of shifts in underlying processes being implemented at those points in time. It is difficult to determine the specific intervention changes from the data and records used in this study, and there are no references to any specific changes in the EAU reports that might have brought around these jumps in outcomes. Furthermore, the analysis of context has suggested a relatively consistent set of test constructs with occasional changes to the cognitive load of test items for different exams at different times. This would lead to an expectation of a more continuous regression rather than discontinuous jumps. The analysis of context does not therefore suggest an explanation for the discontinuity events seen in Figure 9-4.

Although the change in outcomes in 2000 coincides with the introduction of the NMC, the jumps in 2004 and 2008 do not coincide with the introduction of the FACTS in 2005. Borman et al. (2003) have argued that any impact on outcomes would be expected to show a time-lagged jump or gradual improvement over the years following the introduction of policies. However, the discontinuous regression seen in Figure 9-4 supports the possibility of more directed systemic changes to the examination processes and procedures rather than a direct influence by these two policies. It also shows that compared to the pre-NMC decade, which saw one sustained jump in outcomes, the post-NMC decade saw three (2000, 2004, and 2008).

As mentioned earlier in section 3.3, the multidimensional nature of large-scale policy introduction would have influenced different aspects of the educational landscape through different means. These results suggest that the introductions of the policies had an indirect

impact on restructuring the high-stakes examination by affecting more specific policies and procedures directly associated with the JLEE examinations.

9.3.3 Variations in rates of improvement

This section compares the decade before the introduction of NMC to the subsequent decade to support inferences drawn from the analysis of the pass-fail rates in the previous section. It focuses on the differences between %pass and %fail rates (Δ%) to allow a clearer visual interpretation of change. Although doubling the observed change seen on the %pass rate graphs does not add much in terms of measuring the changes in outcomes, it was the opinion of this research that it enhances the visual interpretations of the graphical displays and has been used consistently throughout.

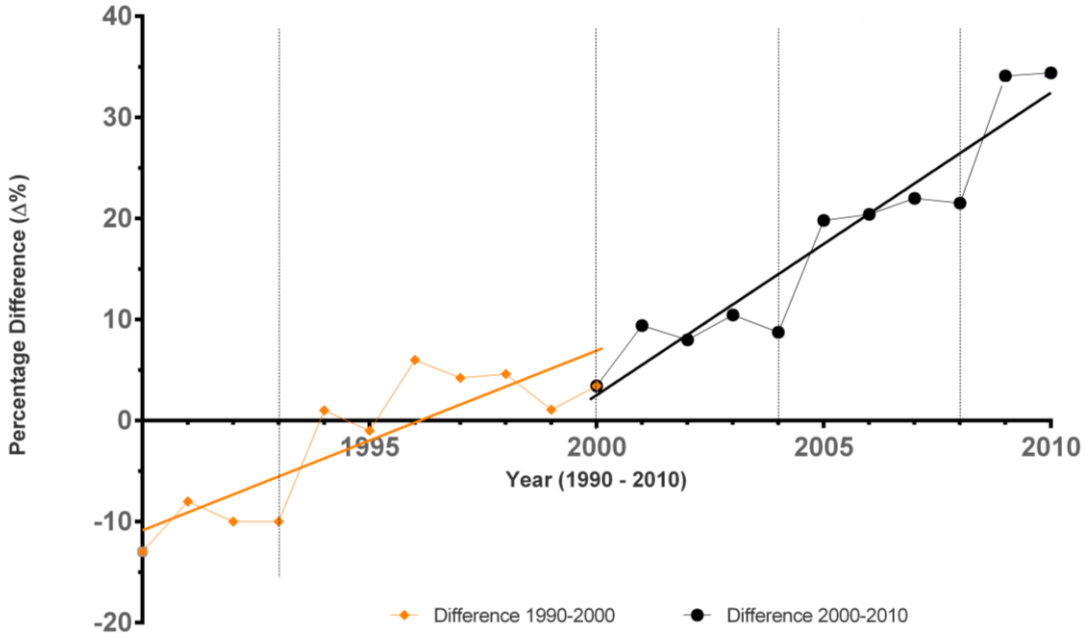
Comparing the regression discontinuity in the Δ% shown in Figure 9-4 shows that there was an overall increase in the rate of change in student outcomes from one decade to the next as reflected in the trend lines. Figure 9-5 shows the plot of Δ% against time in years and shows an increase in the slope during the last ten years when compared to the previous decade. Comparing the gradients of the Δ% for the periods from 1990-2000 and 2000-2010 indicates that there was an increase in the rate of change from one decade to the next.

The trendline gradient for 1990-2000 (determined using GraphPad 6.0 tools) was 1.779 ± 0.3654 %p.a. while that for 2000-2010 was 2.992 ± 0.3299 %p.a., showing an accelerated rate of percentage gain for the decade after NMC compared to the decade prior. Effectively, this represents a difference of 1.21 ± 0.70 % p.a.[6] for the increase in the rate of improvement of outcomes (Δ%) or 0.62 ± 0.43 % p.a.[7] increase in the number of passes between decades.

[6]Uncertainty $\Delta z = |\Delta x| + |\Delta y|$

[7]Uncertainty $\Delta z = z(\frac{\Delta x}{x} + \frac{\Delta y}{y})$

Figure 9-5 Rate of change of percent difference (1990 - 2010)



The increased gradient and associated margins of error suggests an increase in the rate of improvement from one decade to the next and analysis of year-on-year variations indicates that changes took place in stages during the latter decade. The punctuated jumps in outcomes themselves cannot be directly related to the introduction of the two main policies, however, the analysis of context has pointed to efforts to improve the assessments directly by modifying extraneous load. These modifications varied in scope, type, and degree of change for each of the different subjects and took place at different points on the timeline suggesting that this was not a coordinated effort across the five subjects. Similarly, some changes took place to the intrinsic loads embedded in the syllabi.

The general overall trend shows that the rate of improvement increased in the decade following the introduction of the NMC. The effect of the FACTS policy is not so clear as the data only reflects a five-year period, and although there is a jump between 2008 and 2009, there is no information to link it directly to either policy. The jump in 2009 coincides with the introduction of resit exams for those students who failed one subject, however, as shown earlier, the resits themselves did not impact the data used in this analysis. The jump in the pass rate of 6.3% is due to other systemic changes that took place at the same time.

9.4 Aggregated grade average

This section looks to address the skewed outcomes data resulting from the EAU's conjunctive criteria by aggregating the student scores to derive an overall average for each student. The analysis of aggregated grade averages was able to deliver some added insights into gains over

the years and focussed on the variation in percentage difference (Δ%) following the introduction of the NMC in 2000. This was done to establish comparisons on the same scale as the difference in pass-fail rates from the previous section.

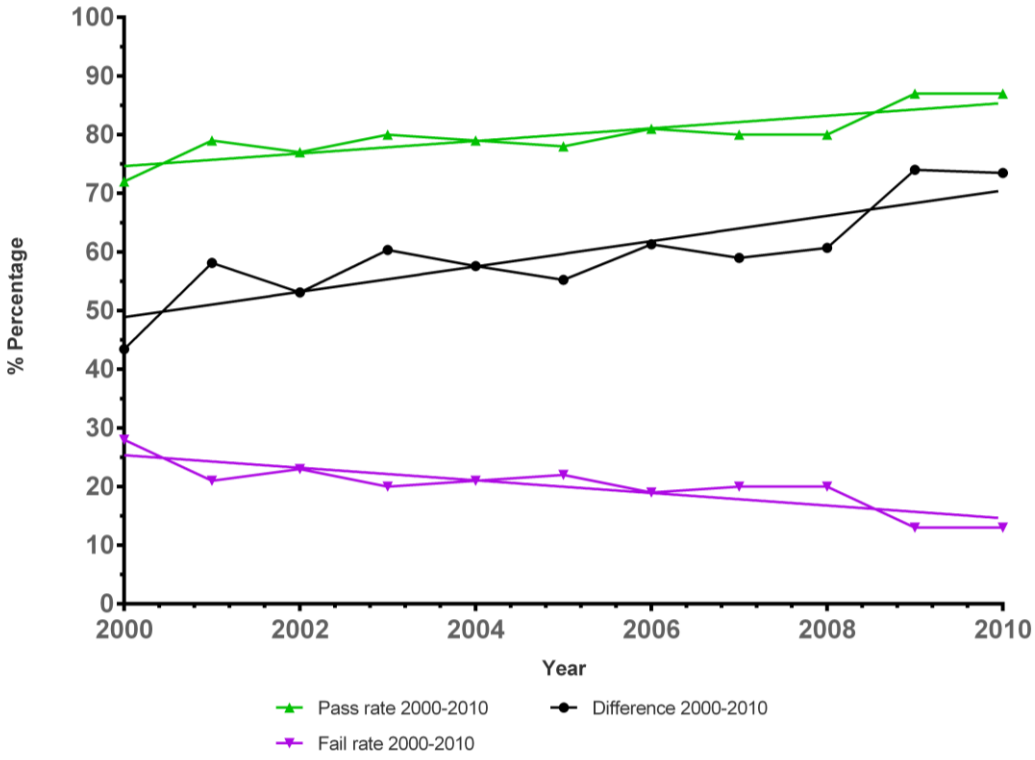
Table 9-4 above is a record of the pass and fail percentages based on the ministry's success criteria, and the differences between these two percentages were analysed longitudinally for variation. As discussed earlier in section 5.4.2.2 (Aggregated grade average) however, the actual measure of achievement was skewed by the EAU's conjunctive criteria (Douglas & Mislevy, 2010; McBee et al. 2014) as students achieving any one grade below C failed the JLEE overall thus affecting achievement ratios. This issue is also noted by Grima et al. (2008, p. 69). The aggregated average for each student was therefore calculated using the procedures described in 5.4.2.2 and used to determine if there was any variation in overall achievement without the bias of the EAU's conjunctive criteria.

However, determination of variation in achievement using this method had a reduced comparative validity due to the norm-referenced system for determining grade boundaries and the fact that this system was not applied consistently by the EAU after 2002. The discussion presented in section 5.4.2.2 though argued that this process could still be contrasted with the outcomes from the previous section, as the pass criteria (C Grade) maintained a consistent 50% cut-off mark. This validity would have been improved had the policy been applied consistently or if the averages were determined using the students' raw scores.

The plots presented below in Figure 9-6 show the *Pass-Fail rates based on aggregated averages (2000 - 2010)*, which offers an analysis of achievement using a compensatory model (Douglas & Mislevy, 2010; McBee et al. 2014) rather than a conjunctive model. For comparative purposes, the colour and symbols used are the same as those used above in Figure 9-4 *Pass-Fail rates (1988 - 2010)* for the same period.

The rate of change of difference (Δ%) as determined from the trendline gradient (Figure 9-6) using GraphPad 6.0 tools was 2.160 ± 0.4723 %p.a. which implies an overall improvement in passing rates of 1.073 ± 0.2279 %p.a. Although the trend shows a similar general improvement in the pass rates over the eleven years to those in Figure 9-4 *Pass-Fail rates (1988 - 2010)* above, some differences stand out in the year-on-year analysis and are considered in the next section.

Figure 9-6 Pass-Fail rates based on aggregated averages (2000 - 2010)

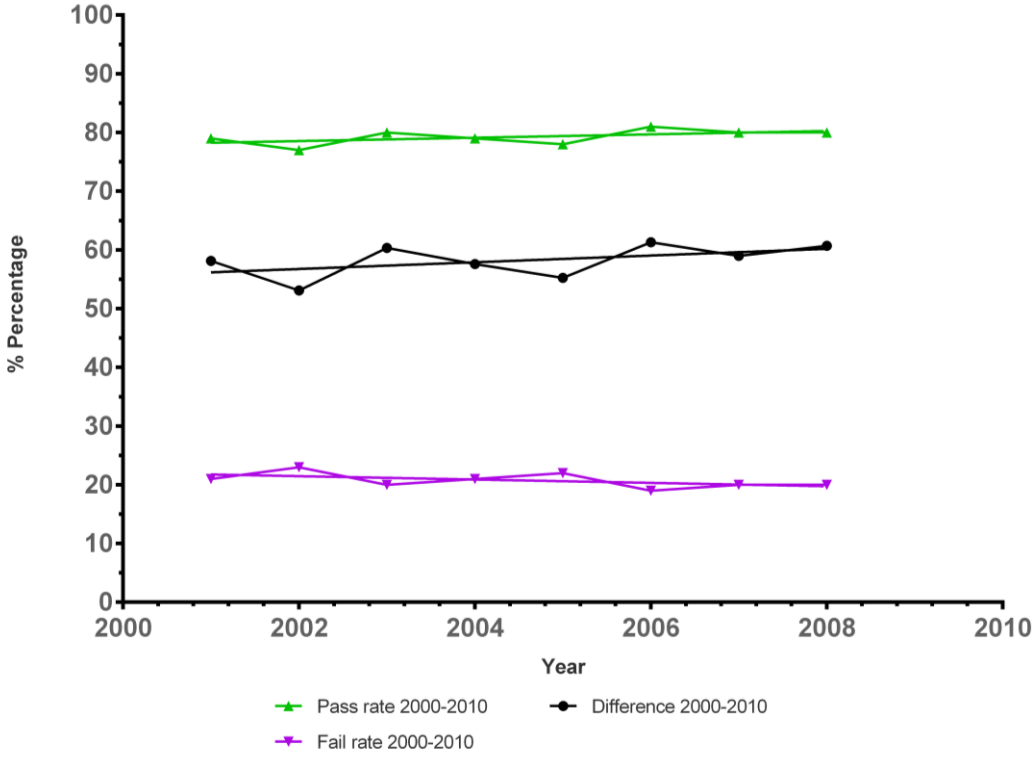


9.4.1 Comparing aggregated grade averages and pass-fail rates

In comparing the trend lines between Figure 9-4 and Figure 9-6 from 2000 to 2010, there are common upward trends in achievement between 2000 and 2001 and again from 2008 to 2009. However, the improvement seen in Figure 9-4 from 2004 to 2005 is not repeated in Figure 9-6.

In actual fact, when the upticks at either end of the horizontal scale (2000-2001 and 2008-2009) are excluded, the plot of aggregated averages actually shows a consistent level of outcomes for the eight years in between as shown in Figure 9-7. The gradient for Δ% over this period is 0.5714 ± 0.4102 %p.a. indicating a yearly improvement in the pass rates of 0.2857 ± 0.1790 %p.a. These are marginal changes and reflect an eight-year period where results and outcomes were fairly consistent. It is notable also that this period spans a change in directorship that influenced other aspects of the test as implied by the item analysis but not on the overall outcomes.

Figure 9-7 Pass-Fail rates based on aggregated averages (2001 - 2008)



9.4.2 Implications of aggregated grade average analysis

Within the context of this investigation, the analysis indicates that there was little to no change in the overall level of student achievement between 2001 and 2008. Furthermore, this suggests that the improvements seen in 2000-2001 and 2008-2009 were likely brought about by specific modifications to the exam setting and marking standards for those years rather than an improvement in achievement levels due to better overall student learning. The discrepancy between the two analytical sets (pass-fail rates and aggregated grade averages) for the 2004 – 2005 period needs further investigation.

The next two analytical sections were not initially forecast in the initial methodology but have subsequently been added to the analysis. Their purpose was to investigate each subject examination more closely and determine if their influence and impact on the overall success rates were cross-sectionally similar or distinct.

9.5 Grade proportion distributions – Banded grade scores

The final analysis of the results focused on the distribution of banded grades for each subject with a longitudinal and cross-sectional investigation of possible trends in grade distribution along and across those subjects.

Due to the occasional norm-referenced determination of grade boundaries, and inconsistent year-on-year criteria (Table 9-6 - % distribution of A and B grades by subject (1997 -2010)), direct longitudinal comparison of the individual grades was not a reliable option for analysis. Rather, a grouping of the passing and failing grades was prepared by combining the numbers for each of the subjects such that a comparison could be made of (A+B+C) vs (D+E). This was established on the pass (C grade) being consistently set at ≥ 50%. Absenteeism and exemptions were not considered as part of this comparative analysis as the purpose was to understand the impact on examination outcomes and achievement.

9.5.1 Shifting grade boundaries

An analysis of the application of the norm-referenced grade boundaries (Table 9-6) shows an inconsistent application of this policy both longitudinally and cross-sectionally.

Table 9-6 - % distribution of A and B grades by subject (1997 -2010)

	Social Studies		Maltese		English		Maths		Religion	
Annum	Grade A (%)	Grade B (%)	Grade A (%)	Grade B (%)	Grade A (%)	Grade B (%)	Grade A (%)	Grade B (%)	Grade A (%)	Grade B (%)
1997	4.5	19.8	4.4	23.1	5.3	20.5	4.4	20.5	6.2	18.5
1998	3.9	22.5	3.7	22.4	3.8	20.2	5.3	21.2	6.0	23.4
1999	6.0	20.8	6.0	20.7	5.5	19.7	5.5	20.3	6.7	27.2
2000	5.7	23.2	6.6	20.4	6.0	20.7	5.8	19.8	9.4	19.8
2001	4.6	21.7	5.7	21.3	5.8	20.1	5.5	21.0	5.0	20.9
2002	12.7	31.8	5.5	20.8	5.1	19.7	11.7	26.1	6.7	21.5
2003	13.2	45.1	10.1	26.9	9.9	16.3	17.5	24.4	34.0	44.8
2004	17.2	34.3	9.5	24.2	9.6	14.1	12.4	25.7	24.1	46.2
2005	21.3	25.0	10.6	18.3	10.3	15.0	10.4	18.5	12.5	21.6
2006	11.0	18.2	14.0	32.5	7.4	15.3	14.9	17.1	20.5	35.3
2007	22.7	25.9	7.3	20.7	5.2	16.4	20.6	22.1	15.8	33.6
2008	20.2	25.8	9.8	26.1	5.2	16.3	20.1	17.9	27.4	36.0
2009	22.6	31.6	12.2	29.5	11.3	24.9	20.4	18.6	28.5	30.8
2010	NA	NA	13.9	30.7	11.6	25.5	19.5	21.3	25.2	34.4

Source: Results register (Educational Assessment Unit, 1997–2010)

The EAU reports mention the application of these grade boundaries for each year between 1997 and 2002 (Curriculum Department & Educational Assessment Unit, 2002, p. 9). The table shows, however, that these criteria were not consistently applied for the 2002 sessions. The direct reference to these boundaries was dropped in the 2003 EAU report and was never applied again for any of the following sessions.

Although inconsistently applied in 2002, and discontinued thereafter, there is no distinct variation in the analysis of achievement (Figure 9-4 *Pass-Fail rates (1988 - 2010)*, p.222 and Figure 9-6 *Pass-Fail rates based on aggregated averages (2000 - 2010)*, p.227) that would

indicate any impactful changes as a result. This indicates that defining norm-referenced grade boundaries for grades "A" and "B" alone did not have any direct influence on student outcomes or associated pass-fail rates.

In order to get a clearer picture of what was happening to the overall achievement from the individual grades for each subject, the study needed to band the passing and failing grades and analyse the trend variations over time and across subjects.

9.5.2 Banded grades analysis – subject based

For this part of the analysis, the % difference between banded grade scores (A+B+C) and (D+E) was analysed to determine the longitudinal trends discussed in the previous section.

The plots below (Figure 9-8 - Figure 9-12) were prepared using Excel and show the trends for each of the subjects from 1997 – 2010.

Figure 9-8 Banded score analysis - Social Studies 1997 - 2010

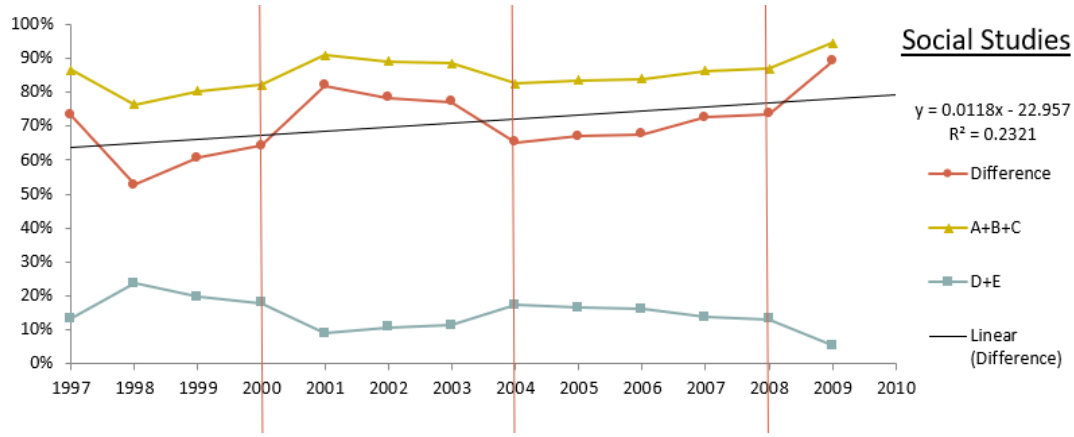


Figure 9-9 Banded score analysis - Maltese 1997 - 2010

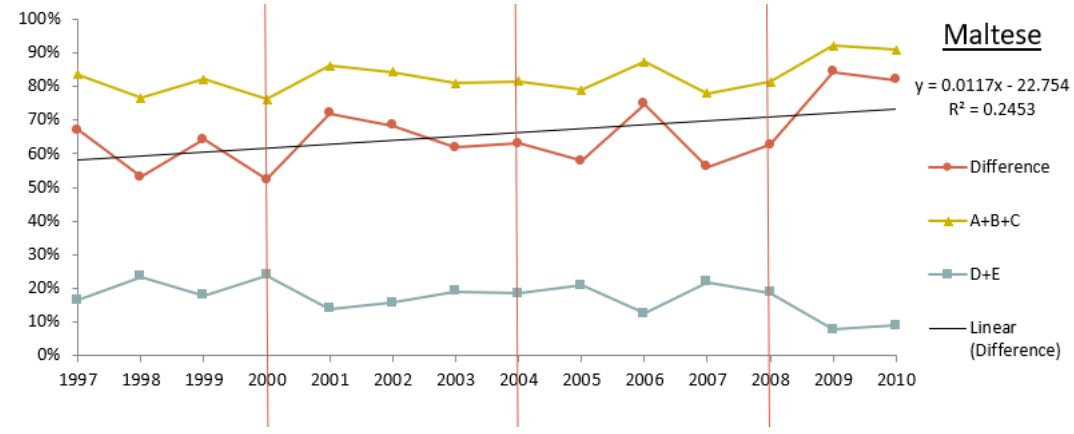


Figure 9-10 Banded score analysis - English 1997 - 2010

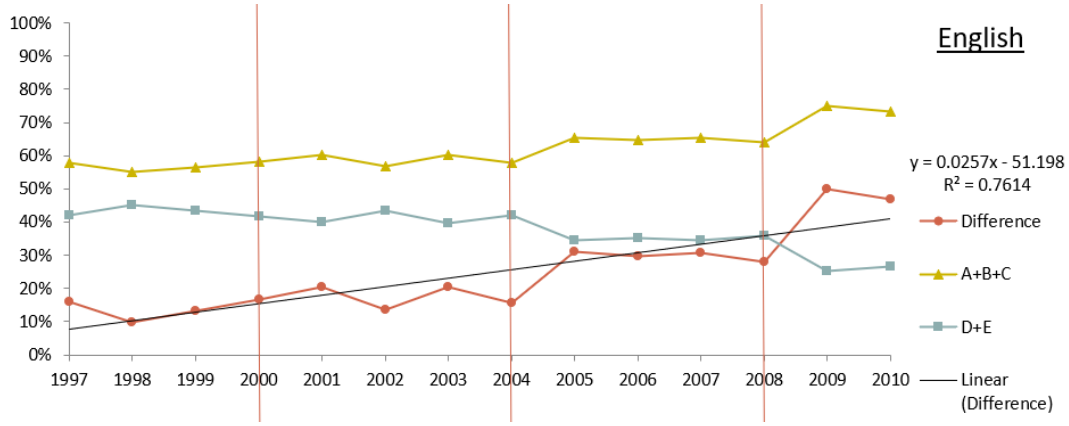


Figure 9-11 Banded score analysis - Mathematics 1997 - 2010

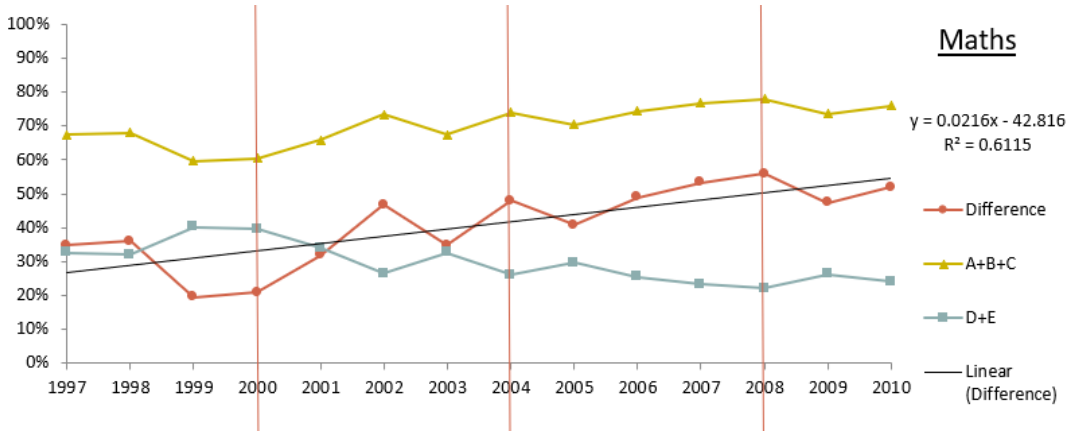
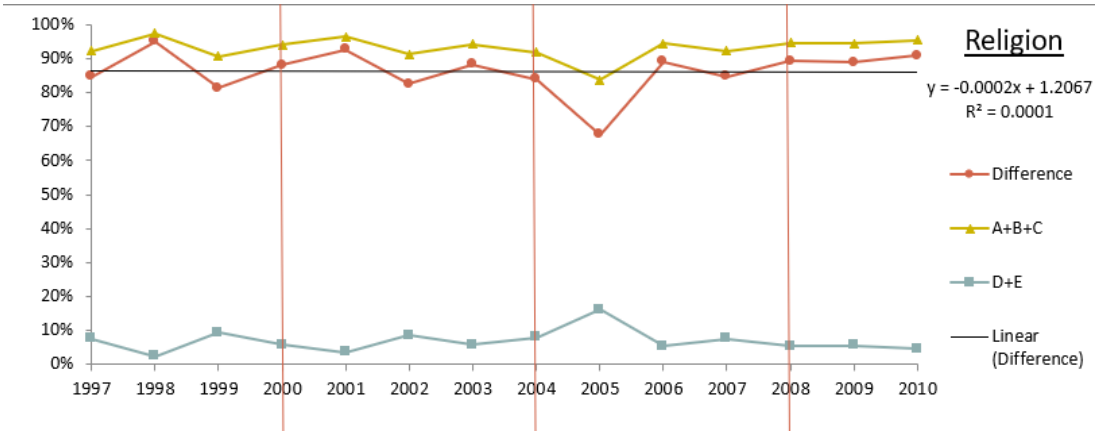


Figure 9-12 Banded score analysis - Religion 1997 - 2010



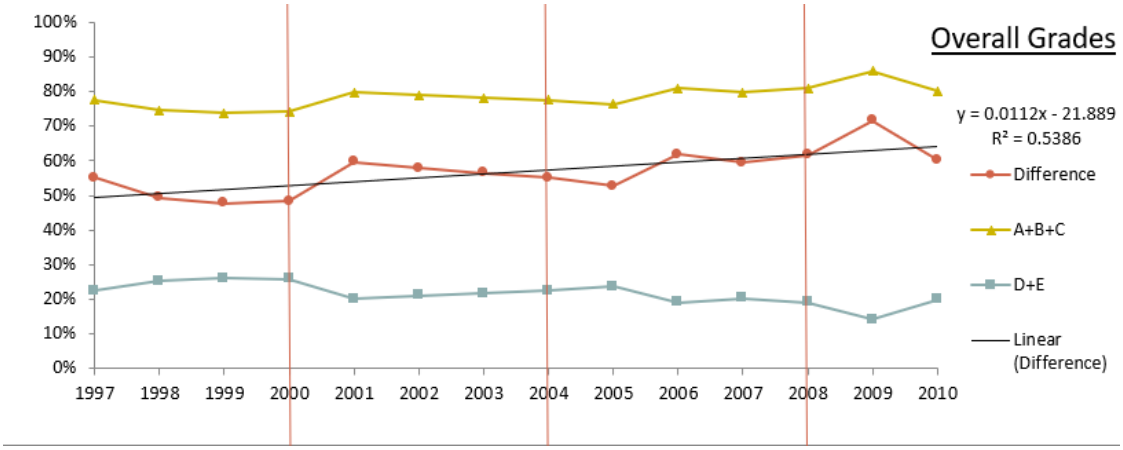
With the exception of religion, there is a general improvement in each of the subjects with the gradient of Δ% of 1.18% p.a. for social studies and 1.17% for Maltese, and a stronger increase observed in English and Mathematics (2.57% and 2.16% respectively). Religion showed little change and maintained a high Δ% (except for 2005) throughout the period 1997 – 2010, indicating very high pass rates for that subject.

This analysis suggests that English had a particularly strong influence on the overall student success rates in passing to the Junior Lyceum. The upward trend from 2004 to 2005 in English (while all other subjects reflect only a slight increase or a drop in Δ%) matches the increase in the pass-fail rates between those years (Figure 9-4 *Pass-Fail rates (1988 - 2010)*). More specifically, from 2004 to 2005 there was an 8% drop in the aggregated grade total (A+B+C) for religion and an 8% increase in the same aggregated grade totals for English. Similarly, mathematics and Maltese dropped by 4% and 3% respectively. However, as English had a greater influence on the rate of success, its impact on the overall pass-fail rates was much more prominent, resulting in the jump observed between 2004 and 2005 in Figure 9-4.

9.5.3 Banded grades analysis – combined

Combining the above plots into an overall achievement graph shows a tighter bandwidth (smaller standard deviation – 6.4%) than each of the individual subjects. (9.6%, 9.9%, 12.3%, 11.5%, and 6.6% respectively). A similar phenomenon is argued by Newton (2021), claiming that aggregating different plots of pass-fail rates for different subjects tend to iron out lower level fluctuations leading to a more consistent year-on-year plot.

Figure 9-13 Banded grades analysis – Combined



This plot matches the analysis of aggregated grade averages showing jumps in achievement in 2000 and 2008 that match the pass-fail rates (Figure 9-4 *Pass-Fail rates (1988 - 2010)*) analysis, but no similar jump from 2004 to 2005.

9.6 Single and combined subject failures

In looking at single subject failures, the study used Excel to aggregate and count those cases that resulted in a student being unsuccessful due to having failed one subject alone. The previous sections showed the dominant influence of English and mathematics on pass-fail rates on the JLEE. This section looks to quantify that influence.

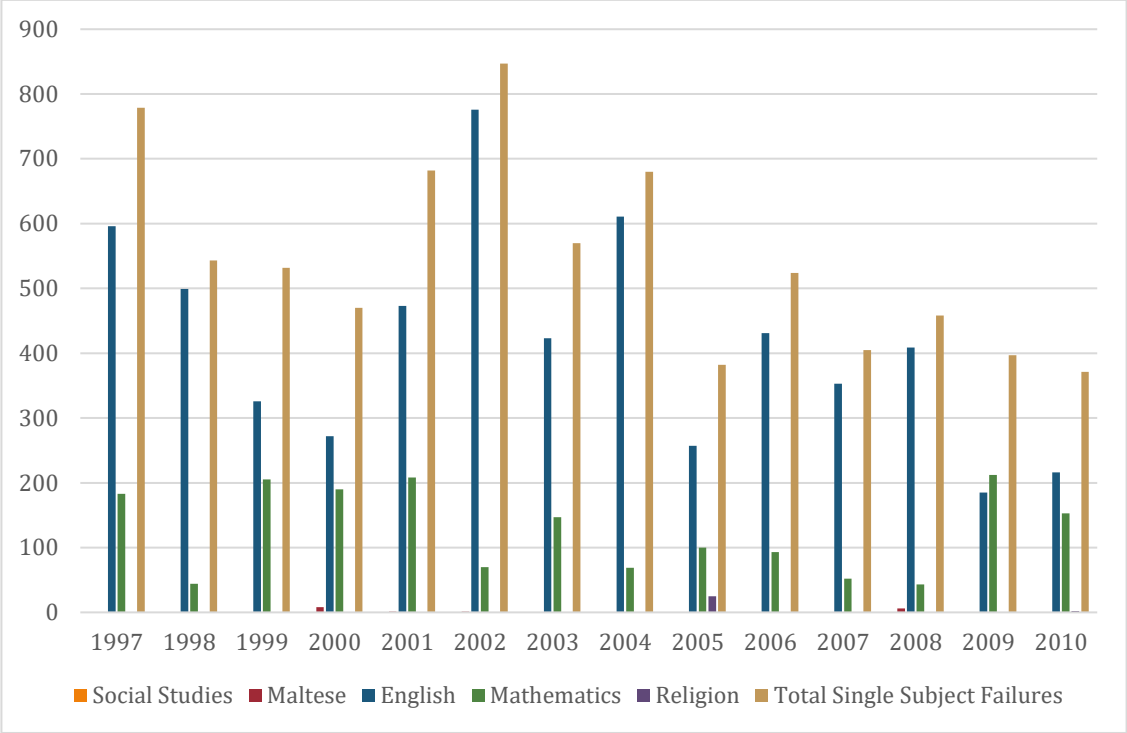
Table 9-7 below shows the number of single subject failures per annum by subject. The corresponding bar chart (Figure 9-14) shows the disproportionate impact of English, and to a lesser extent Mathematics, on overall outcomes. The other three subjects had occasional single-subject failures associated with them, but these seem to be more of an anomaly to the trend.

Table 9-7 Counts of single subject failures (1997 - 2010)

Year	Social Studies	Maltese	English	Maths	Religion	Annual Totals	% of total applicants
1997	0	0	596	183	0	779	17%
1998	0	0	499	44	0	543	12%
1999	1	0	326	205	0	532	11%
2000	0	8	272	190	0	470	10%
2001	0	1	473	208	0	682	15%
2002	0	1	776	70	0	847	19%
2003	0	0	423	147	0	570	13%
2004	0	0	611	69	0	680	16%
2005	0	0	257	100	25	382	9%
2006	0	0	431	93	0	524	13%
2007	0	0	353	52	0	405	11%
2008	0	6	409	43	0	458	13%
2009	0	0	185	212	0	397	12%
2010	NA	0	216	153	2	371	12%
Subject Totals	1	16	5827	1769	27		

The number of single subject failures as a proportional percentage of total applicants indicates relatively consistent numbers over the years.

Figure 9-14 Single subject failures by subject



A further analysis was conducted to determine cases where multiple failures in different combinations of social studies, Maltese, and religion might have played a role in affecting the rate of achievements. The values are presented in Table 9-8.

Table 9-8 Combined subject failures

Year	SS&ML	SS&RL	ML&RL	SS&ML&RL	EN&MT
1997	2	0	1	2	480
1998	1	0	0	0	305
1999	1	0	1	0	615
2000	2	0	0	0	494
2001	0	1	0	0	656
2002	0	1	2	1	392
2003	3	1	0	1	480
2004	5	0	2	3	277
2005	3	0	3	6	262
2006	3	0	1	0	254
2007	6	0	2	7	149
2008	3	0	1	1	183
2009	0	1	1	1	319
2010	NA	NA	0	NA	307

Similar to Table 9-7, the data presented in Table 9-8 emphasises the disproportionate impact of English and mathematics on the general outcomes and achievement of students sitting the JLEE. This corroborates the same understanding drawn from the previous section 9.5.2 Banded grades analysis – subject based. The data also shows a relatively consistent pattern over the 14 years in question and suggests a possibility that this might have been done by

design. That being the case, the implications from these different analyses highlight that the key determinant of successfully passing the JLEE was successfully passing English and Mathematics.

There are no indications that there were any policies proposing a weighted influence of the different subjects on student success rates, nor is there any indication of how this disproportionate impact came about. Whether or not such a design was integrated into the examination papers themselves or the marking schemes is not clear, but social studies, Maltese, and religion had a lesser impact on student success rates than English and mathematics.

9.7 Summary: Outputs analysis

In response to the third research question, the analysis of outcomes was developed to look at variations in performance as a reflection of the quality of education (OECD, 2013). The continuous series of JLEE student examinations was an ideal source of data for understanding longitudinal trend patterns and facilitating this analysis set.

In summary, the analysis of the output data has highlighted five key outcomes:

i. An overall improvement in student achievement

The set of four analytical tools developed to investigate the variations in student achievement and link to the broad-scale introduction of the two national educational policies have shown definitive upward trends in overall student achievement that was sustained in the long term. The overall rate of improvement in student success rates can be observed before and after the introduction of the NMC and FACTS policies in the overall pass-fail rates, but also in the aggregated grade average and banded scores analysis for the individual subjects. What was also noted was that the rate of improvement increased after the introduction of the NMC and FACTS policies.

ii. An increased rate of improvement after NMC

There was a boost to the rate of change in achievement of 0.62% p.a. when comparing the decade before and after the introduction of the NMC. The increase was mainly due to three big jumps in success rates in 2001, 2005, and 2009, compared to only one similar jump in the decade prior to the introduction of the NMC. This increased rate can be associated with the general positive pressures brought about by the broad scale policies influencing different systems across the educational landscape. However, the jump in achievement in 2005 resulted from a particular set of circumstances associated with the English exam in particular.

iii. The variation in outcomes was intermittent

The main changes affecting the increases stated in (ii) took place at the beginning and end of the latter decade with the period between (2001 – 2008), showing relative consistency when using a compensatory model to determine success rates. Coupled with the relative consistency and continuity of the test constructs and forms, this suggests a likelihood that variation in exam moderating standards and implementation policies carried a greater influence on the observed jumps in achievement in 2001 and 2009 than any grassroots variation in teaching and learning standards.

This last argument is given greater weight once the disproportionate influence of English and mathematics on overall achievement are factored in. The uptick between 2004 and 2005 was mainly a result of a greater proportion of students passing English rather than a generally positive change in all subjects.

iv. The impact of applying conjunctive criteria

The four different systems for analysing the data highlighted discrepancies in their implications that could be linked to the application of conjunctive criteria used to determine success or failure. English, and to a lesser degree mathematics, had a disproportionately higher impact on student success rates when compared to the other three subjects.

The applied conjunctive criteria affected success for each yearly sitting. This is borne out by the analysis of data using a compensatory model for deciding success rates. This latter model exposes the relative consistency and only slight improvement in success rates between 2001 and 2008. More specifically, the uptick in the pass-fail rate analysis between 2004 and 2005, seen in Figure 9-4 and Figure 9-5, is not reflected in the subsequent analysis of grade score averages (Figure 9-6) and banded grade scores (Figure 9-13) for the same period. Quite the contrary, when the data was reprocessed to reflect the scores as a compensatory set of outcomes —the aggregated grade averages and the combined banded scores — the analysis of difference showed a slight drop in achievement during that time. The conjunctive showed an increase of 5.6% in pass rates while the compensatory model showed a decrease of -2.4%.

v. The importance of construct continuity

To reinforce this issue highlighted in (iv) above, the construct analysis has shown that there were changes to the statistical score distribution for English in 2002 and 2004. These changes seem to have led to a drop in achievement for English for those years. Once the statistical score distributions were readjusted in 2003 and 2005 back to the norm, the score for English increased again to previous levels and the overall achievement results also returned to their

previous norms. This highlighted the impact of adjustment of the constructs on outcomes and, due to the direct influence of English and mathematics on overall student success rates, the importance of sustaining construct continuity for these two subjects.

The next chapter will discuss the various outcomes that constituted the three main chapters of this analytical section pulling together the general understandings of all three within the context of the quality framework to establish an overview of policy impact on student achievement.

Section 4: Discussion and Conclusions

Section Overview

This section presents the key findings drawn from the analysis and uses the quality framework to make connections across the different domains before discussing associated implications.

It is structured into a single chapter presenting a brief review of the main scope of the research followed by a discussion combining the outcomes from all three analytical chapters and drawing both general and specific insights from the analysis and results section.

10 Discussion

10.1 Introduction

This study of policy impact was motivated by an existing gap in related research following the introduction of the NMC and FACTS policies in Malta. It was guided by a single overarching RQ to determine the tangible effects of policy change on learning outcomes and was informed by a longitudinal analysis of student achievement before and after the policy introductions.

The annual Junior Lyceum Entrance Exams presented a continuous set of data that could be used to underpin the analysis and inform the research. This set of five examinations was taken by the majority of Maltese students transitioning from primary to secondary schooling and is considered to be a benchmark. The data associated with these five exams also offered the possibility to identify cross-sectional trends, nuances, and variations that would indicate broader affecting factors impacting the whole of the examination landscape.

However, the research also needed to confirm an acceptable degree of comparative validity of the examination results over the period being studied (1997 – 2010). This required other factors affecting construct continuity and cognitive loads to be considered. To this end, the longitudinal analysis had four major investigative threads associated with 1) continuity of test constructs 2) consistency of test forms, 3) variation in the mental load of the examinations, and 4) general trends in psychometric characteristics.

With these requirements in mind, the overarching research question was organised into three sub-questions. The first considered the processing methods needed to aggregate and prepare the JLEE data and was partly guided by the nature of the data available, and partly by the expected analytical methods. The collection of this data as hard copy prints, and conversion to digital soft copy, required a dedicated set of structures and procedures to ensure the quality

and reliability of the conversion. The second and third sub-questions structured the work associated with the four investigative threads mentioned above and the overall impact on achievement. These formed the core of the research to investigate the possible impact on examination structures, variations in attainment, and changes in the quality of education reflected in achievement scores.

In considering the general framework of the research, it became evident early on that any impact would be multivariate and cascading across different educational domains. The study was therefore underpinned by a structured investigation based on the quality framework presented by Scheerens et al. (2011a) classifying various educational functions under four principal domains: *Input – Process – Output – Context*. This framework had been applied by various UNESCO reports regarding the monitoring of quality education (UNESCO, 2002, 2005), albeit for different purposes. Although considered narrow and linear in nature (Scheerens et al. 2011a), this framework offered connectivity across the four domains that allowed the research to categorise the analysis for those connections. Furthermore, this framework simplified the organisational structures and sequencing of a multidimensional investigation and restrained the complexity of analytical processes by allowing singular pathways of investigation that could be connected longitudinally and cross-sectionally (Figure 5-4 Outcomes analysis flow diagram).

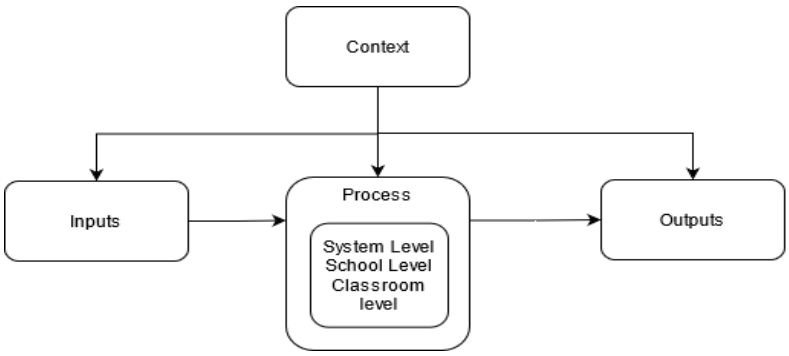
10.2 Key Findings

The study demonstrated that the overall rate of improvement in student outcomes on the JLEE was consistent but not continuous after the introduction of the NMC and FACTS policies reflecting intermittent, systemic changes in the implementation procedures for the exams. However, these progressions could not be directly linked to the policy documents themselves. The structuring of the policies, although replete with progressive intentions, lacked purposeful, overarching mechanisms that could be applied to guide the implementation. These mechanisms were later integrated into the next broad-scale policy (MEDE, 2012), and included performance indicators without which it is impossible to benchmark success (Lennon, 2016, p. 154).

The study also determined that the policies had little impactful effect on teaching and learning beyond what is generally improving results over time due to maturing classroom instructions. Such policy needs to have a meaningful impact on teaching and learning in very explicit ways to have some form of measurable summative effect on benchmark examinations (Sackney, 2007).

In considering the quality framework used for the investigations (outlined in Figure 10-1 below), the study found this system to be an effective mechanism that allowed a structured analysis of each of the i-p-o-c domains and their subsequent linking. The findings for each domain are reviewed briefly before presenting a more detailed discussion and establishing connections in the following sections.

Figure 10-1: Quality Framework - Scheerens et al. (2011a, p. 36)



Although offering connective relationships across the different domains, some variations to Scheerens' model were made for this particular study. More specifically, the adopted framework integrated the analysis of context with the input and process domains but not with the output domain. The latter integration would have had to consider germane factors and characteristics associated with the subsequent year groups and this, as argued earlier, was not feasible.

Before presenting more detailed discussions and associated implications in the following sections, the general determinations for each domain can be briefly recapped as follows:

i. The inputs and context analysis showed that consecutive policies underpinned a continuous, evolving process to *improve the development, implementation, and monitoring systems for enhancing the quality of education.* However, monitoring and feedback systems were not effectively integrated to support implementation.

ii. Process and context analysis determined that although the general constructs remained relatively consistent over the fourteen years, there were observable changes in the contextual underpinnings. The five different examinations showed different procedural approaches when it came to exam structures and settings. As a result, construct and form changes exhibited variations in properties at different times.

Furthermore, analysis of the psychometric characteristics showed that all five exams became relatively easier over the fourteen years although the dimension of quality, as reflected by the discrimination index, remained fairly consistent.

iii. The outcomes analysis indicated that although there was a sustained rate of improvement in achievement scores, the main drivers were associated with systemic changes influencing the conjunctive criteria. More specifically, this was a direct consequence of the English and mathematics results on student outcomes. This suggests that it would not be possible to make determinations about any policy impacts on the effectiveness of teaching and learning at the grassroots level through the analysis of the outcomes alone.

10.3 Implications of findings

This section begins with a brief outline of what are considered to be the primary implications of this study before reviewing the key findings in further detail within the context of each of the quality domains. The information is then triangulated to outline the connections across the different domains that led to the drawing of these implications.

10.3.1 Primary implications:

i. The broad-scale policies working across the Maltese educational system needed evaluative tools to understand and guide implementation and subsequent impact. The policies also needed more specific performance indicators associated with learning outcomes to measure progress.

ii. The directly controlled modification of background assessment processes had a greater effect on the JLEE outcomes than the implementation of the two policies and was likely used to influence pass rates between 1997 and 2010.

iii. The improvements in achievement levels between 2000 and 2010 were due to direct adjustment of English and mathematics exam processes rather than an influence on grassroots teaching and learning processes.

iv. Considering the JLEE was intended to be a broad measure of student achievement across five subject areas, English and mathematics had the greater influence on outcomes due to the EAU's conjunctive criteria to the extent that the other three subjects were practically irrelevant to overall success rates.

v. The changes in the construct of the English exam in 2002 and 2004 emphasise the importance of maintaining construct continuity to ensure monitoring validity during the implementation process.

10.3.2 Inputs and context: Policy analysis

The policy analysis confirmed the intention of the authorities to sustain continuous improvement in the quality of education (MEYE, 1999, 2004a), however, it also highlighted a

paucity of predefined systems to monitor that quality. The NMC and FACTS recognised the general principles for evidence-based decision-making to guide implementation (Galea, 1999, 2004; Mizzi, 1999) but did not attempt to structure such systems through the policies themselves. By contrast, the subsequent broad-scale National Curriculum Framework for all (MEDE, 2012) integrated a thorough monitoring system in its policy structures.

Consequently, although the NMC and FACTS policies led to concrete actions on the part of different stakeholders and likely influenced the observed increase in rates of improvement, any effects on student outcomes could not be directly traced to any particular part of the implementation. The NMC and FACTS created a permeating, constructive reform pressure influencing the whole of the educational system and the associated analysis of results offered supportive arguments that this positive pressure affected student outcomes. However, the shortcomings in oversight and effectiveness research meant this impact could only be associated with policy action's general stimulus and a more direct linkage could not be made.

The evidence supports arguments that these large-scale initiatives drove change at subordinate tiers and also created additional levels. Different educational areas applied reforms according to their own contexts with the establishment of new directorates being arguably one of the more evident changes impacting the system. This meant that although dedicated quality controls were not explicitly stipulated, the policies had, for example, set the scene for establishing the Directorate for Quality and Standards in Education. The subsequent introduction of "Knowing Our Schools" (MEYE, 2004b) went some way to bridging a gap in quality oversight, defining quality controls, attainment measures, and KPIs as reflective measures of quality. However, this was done in a separate manner distinct from the original policy (Section 7.4) and more specific questions associated with affected systems of learning and their impact on achievement and attainment remained ambiguous.

Despite the ambiguity, the analysis of results of the JLEE reflected adjustments in subject-based processes suggesting different areas making changes at different times. However, there were also similar trend variations in the five sets of D_i vs. F plots pointing to a higher tier influence of the examination processes as a whole. There were, therefore, multiple dimensions of influence affecting changes throughout the examination system which likely stemmed from the overall drive for change instigated by the policies.

These findings have important implications for the future development of large-scale policies and highlight the need to establish systems that monitor, investigate and evaluate the effect of policy introduction at different levels of the educational system during the implementation

phase. Although the analysis showed that establishing integrated mechanisms to monitor ongoing progress or effectiveness was left to the individual schools or colleges to consider, there are no reports or evidence to support such actions taking place. Furthermore, the organisational power of centralised administration, as is the case in Malta and Gozo, enables more effective measures and investigations of such policy rollouts and is better suited to support proactive or reactive adjustments to be taken accordingly.

10.3.3 Process and context: continuity and consistency

In responding to the second research question and as part of the quality framework structuring the study, the process and context chapter presented a multidimensional analysis of the examination sets. This investigated the continuity of intrinsic factors and consistency of extraneous characteristics affecting the mental load of the exams before investigating variations in outcomes. The evidence presented in the analysis chapter informed an understanding of the different variables influencing the outcomes of the JLEE.

10.3.3.1 Intrinsic CL

In assessing the intrinsic CL, the study compared test construct validity by looking at the long-term alignment of content structures and statistical specifications. The results indicated that, for the most part, the constructs remained relatively continuous for social studies, Maltese and English and changed in part for mathematics in 2007. Continuity for religion was indeterminate. Construct consistency, however, varied slightly for social studies and mathematics, saw a two-year variation in marking distribution for English in 2002 and 2004, and was indeterminate for Maltese and religion.

10.3.3.2 Extraneous CL

In considering the extraneous CLs, the analysis was more complex due to a multiplicity of affecting influences. To investigate variations in the exams, the analysis drew on work by Newman et al. (1988) to structure both an analysis of the cognitive and statistical characteristics of the test forms as well as psychometric traits. The first part of the analysis compared readability levels, cognitive item demands and general formats and structures of the items, while the second analysed longitudinal variation in facility and discrimination indices. The latter analysis also highlighted changes in the quality of parallel test forms and any associated variation in standards.

Cognitive demands varied slightly over the years and showed no indication of simultaneous cross-sectional changes across subject areas. Rather, the piecemeal variations for the different subjects reflected disconnected changes. However, in considering the psychometric

characteristics, there were cross-sectional similarities in the graphical arrays that indicated higher tier influences caused the exam outcomes to shift in tandem. Furthermore, although the exams became relatively easier overall (as reflected in the analysis of the statistical mean of F and pattern shifts in the D_i vs F arrays), achievement from 2001 to 2008 based on aggregated averages remained fairly consistent. This was supported by the statistical analysis of the discrimination index indicating that the quality of all exams retained consistency during this time except for three years from 2001 – 2003.

There was a notable change in the distribution pattern on the D_i vs F plots after 2003 from approximating an ideal test curve to a more scattered distribution. This showed a more random allocation of good and bad quality items while sustaining the same overall level of quality, implying that how the exams were processed changed without affecting the outcome characteristics of the success rates.

10.3.3.3 Influence of the English Exam

The analysis of the English exams was particular in highlighting the influence of conjunctive criteria determining success and the importance of sustaining longitudinal construct validity of benchmark examinations.

There were variations in the construct for English in 2002 and 2004 that resulted in a detectable impact on overall pass rates in English and caused a drop in JLEE success rates for those two years (Figure 9-10 Banded score analysis - English 1997 - 2010). These variations redistributed score weightings from the comprehension section to the language and grammar sections and resulted in a drop in the proportion of students that passed English. Furthermore, although the construct changed for those two years, the readability of the comprehension texts remained consistent from 2002 to 2004 and became more difficult in 2005 (Figure 8-6 and Figure 8-7). This suggested that the comprehension section had little influence on the outcomes in English.

More notably, however, the changes to the construct impacted the validity argument for the English exam, subsequently affecting any longitudinal comparison with those two years. This underscores the importance of maintaining continuity in the construct definition not only to sustain continuous testing standards but also to support impact and monitoring processes.

Considering student success rates on the JLEE, the study found that there was a disproportionate influence of the English results on the proportion of passes and failures over the years. A similar influence was linked to the mathematics exam but not to the other three exams. This indicated that controlled changes to the construct or exam processing led to

changes in overall outcomes and affected the number of successful candidates shown in Figure 9-4 and Figure 9-5. Consequently, the JLEE had a singular means of adjusting pass-fail rates based solely on the English and mathematics exams allowing such adjustments to take place to support ulterior purposes.

Due to this specific ability to manipulate success rates and the actual changes seen on the pass-fail plots, it is impossible to determine the true impact of policy on students' learning. There is no objective measure in this regard that can be drawn from the JLEE. Such exams need to be set against pre-established standardised processes and allowed to work within those frameworks to effectively determine insights due to introduced policies or processes. The ease with which English and mathematics could be adjusted and the consequences of such actions on success rates, reduce the reliability of such benchmarks to reflect the true state of teaching and learning at that level.

10.3.3.4 Psychometric analysis and the quality of the JLEE exams

The similar cross-sectional fluctuations in the psychometric characteristics over the twelve years (1999 – 2010) suggested possible post-examination processes influenced the final results and also reflected gradual changes to the quality of the different exams.

An initial hypothesis of the research, drawn from the literature review, linked the discrimination and facility indices to the quality of tests and test items. The shifting patterns shown on the D_i vs F plots were associated with the reduction in quality determined by the cognitive analysis of the test forms. The variations in both the cognitive and psychometric analysis were gradual but moved in tandem. The change was more pronounced in the social studies and religion exams, where the cognitive analysis showed a change in assessment tools that led to a reduced capacity to measure higher cognitive abilities. This was associated with a drop in test difficulty levels, making the assessment more lenient in its purpose and suggesting social studies and religion were given less weighted importance as benchmark measures, despite being included with the set of five JLEE exams.

In the context of this study, a greater proportion of questions with a lower cognitive demand were therefore associated with a shift in the plot densities towards the lower right quadrant of the graph and indicated a drop in examination quality. This suggests the study supports the initial hypothesis linking mentioned above, linking D_i vs F plots to the quality of the exams. It also suggests a possible use of D_i vs F plot arrays to help moderate examination quality in the medium to long term for such benchmark examinations.

10.3.3.5 Readability of Maltese texts

There is an inadequacy of research into readability of Maltese language texts that can give a reliable measure of mental load exerted by a Maltese text. However, one other key outcome drawn from this study relates to the readability of Maltese as a non-English text. The analysis of the readability of the Maltese texts demonstrated that LIX, ARI and CLI algorithms can be used effectively to investigate the difficulty of these texts with each showing a similar pattern of variation for the same texts.

10.3.4 Outputs: Outcomes and achievement analysis

The analysis of outcomes and achievement were key analytical components of the output domain within the quality framework, structured to respond to the third RQ and purposed to identify impact as longitudinal variations in outputs. The longitudinal analysis of student results from the JLEE was used to measure any change in education quality reflected in both attainment and achievement over time and shows a general improvement in the overall level of outcomes for the period in question.

The study adopted a "Black Box analysis" (Cohen & Hill, 2001), to determine general effects acting at a macro-level (Section 3.3) and analyse the longitudinal variations in achievement and outcomes. The three key analytical actions — variation in pass-fail rates; aggregated grade averages and; proportional grade distributions on standardised tests — showed an improvement in achievement rates. However, when linked to the process section of the analysis, these improvements were mainly a result of adjustments to the examination processes rather than an improvement of quality in schools.

Another significant aspect associated with the comparison of pass-fail rates showed a more pronounced increase in the rate of improvement in the decade following the introduction of the NMC suggesting an overall impact resulting after the policy introductions. However, this direct tracking of pass-fail rates masked the multi-dimensional complexity associated with the administration of the five subject examinations and the effect of conjunctive success criteria. The improvement observed was characterised by staggered jumps in the pass rates (2000 – 2001 and 2008 – 2009), indicating direct systemic changes to the exams rather than gradual improvement in learning at the grassroots level. As stated earlier, the analysis of aggregated grade averages revealed that the average grade scores remained relatively consistent between 2001 and 2008 with little actual improvement during those seven years.

This outcome was also supported by the combined banded grade score analysis (Section 9.5.3) which exposed the predominant influence of English, and to a lesser extent mathematics, in

affecting the pass-fail rates due to the conjunctive criteria. A notable jump in student pass-fail rates between 2004 and 2005 was due to a change in English grades, while at the same time, other subject grades showed an actual drop in overall achievement.

The analysis of grade proportion distributions provides further evidence that student success rates were almost entirely dependent on passing English and mathematics when cross-referenced with the pass-fail rates. Analysis of the combined subject failures (Table 9-8) emphasises this point further, showing that throughout the years single and combined failures in social studies, Maltese and religion had a minimal impact on the overall outcome statistics of the Junior Lyceum Entrance Exams.

The JLEE could, in essence, have been reduced to an English and mathematics exam and the outcome results would have only shown a slight difference in the pass-fail rates compared to the full set of data. From an application point of view, such a reduction would have inevitably reduced the emphasis on the other subjects delivered in schools by reducing their perceived importance and this may have been an undesired result thus leaving the three exams in place.

It cannot, therefore, be determined if the NMC and FACTS policies had any sort of influence on educational aspects associated with social studies, Maltese, and religion. On the other hand, the variations in English and mathematics were due to adjustments made to the examination constructs and implementation structures of the exams and subsequently affected the attainment rates. This implies that within the context of delivering the NMC and FACTS policies, attainment was strongly influenced by adjustments to the examination structures and the actual impact on the teaching and learning of the students remains unclear and unmeasurable using this set of data.

Although the influence of the English and mathematics exams on overall success rates was evident, the analysis of banded grade scores demonstrated a slight overall improvement in all subjects except religion over the fourteen years. The improvement in English and mathematics was double that of social studies and Maltese and can be attributed to English and mathematics results starting at a lower percentage achievement rate and subsequently making greater gains in response to actions to change the educational systems.

10.4 Connecting domains.

The study, established by the quality education framework (i-p-o-c), effectively underpinned the inquiry, and structured an environment in which to connect policy changes to the analysis of data and variations in outputs, supporting causal associations or connections. This process

determined that although policy did not have a direct impact on specific educational processes, other external influences did.

Comparing the input and process domains longitudinally led to an understanding that there were no direct or immediate influences of policy on any of the constructs, test forms, cognitive structures, or statistical characteristics as a direct result of the policies. Some changes that worked to reduce the extraneous load took place gradually and at different points on the timeline for different subjects. The randomness of these variations in construct and cognitive characteristics for the different subjects indicated that any influences from the policy introductions were not uniform. Had the policies imposed a more immediate influence on the examination structures, there would have been some form of commonality in the trend patterns reflected in the analysis of constructs, forms, and cognitive loads. Combined with the analysis of aggregated grade averages, the study determined that each subject area implemented changes to their processes independently of each other. These observations confirm Grima et al.'s, (2008) assertion that modernising the syllabus had a positive impact on examination design. It also gives weight to prior arguments that the policies had a pervasive positive influence across different educational areas at different times and to varying degrees.

Although not showing any anomalies that might be linked to the policy introductions, the psychometric analysis indicated a strong cross-sectional similarity in trends from 1998 to 2010. One broad-scale variation that could be seen across all five subjects in a consistent and comparable manner was the statistical change to the psychometric characteristics between 2003 and 2004. This has been associated with a change in the EAU director at the time and supports the hypothesis that there were uniform systemic changes in higher-tier examination processes. It also lends further support to the argument that the JLEE exams were easily adjusted according to exigencies and could not present an objective standard against which to compare progress in teaching and learning in schools and classrooms.

Similarly, when the input and output domains were compared, there were no direct associations that could be established between the policies and overall student achievement. The 3% jump in success rates between 2000 to 2001, happened at the same time that the NMC was introduced. If the policy impact was related to changes in teaching and learning, such improvements would be expected to exhibit some degree of time lag following the policy introductions and a more gradual rate of improvement. It would be understandable, however, that following the launch of the NMC, various departments would have begun to make changes with the EAU following a similar trend to implement new procedures resulting in a jump in achievement rates.

The final associative consideration between process and output was more notable than the previous two. The greater degree of influence determined between the process and output domains suggested adjustment of the examination structures and processes, rather than improvement in teaching and learning outcomes. The long-term continuity of the JLEE constructs enabled the longitudinal comparison of outcomes to be conducted. There were occasional variations to the constructs in English and mathematics with no noticeable effects on overall improvement. The analysis of those outcomes indicated that there was no immediate, time-lagged, or direct impact of the NMC or FACTS on achievement or attainment in the JLEE. Rather, variations in achievement seem to have taken place intermittently without a regular pattern that could be associated with the introduction of the policies or the changes in those constructs.

This intermittent and unassociated variation supports the argument that the permeating nature of the policies across the different educational dimensions led to initiatives within or across the various tiers that precipitated affecting actions. It is possible to argue that the actions may have taken place independently as seen in 1994 – 1995, however, the jumps shown on the longitudinal analysis of achievement became more frequent during a period of rapid change brought about by the new policies.

10.5 Summary - Overall impact

This section is structured to present a collective summary of the various determinations and implications discussed in the last two sections.

The main understanding determined from this research supports the argument that there were indirect effects brought about by the implementation of the two key policies, creating a progressive environment across the educational landscape. This created a developmental pressure on quality across the various educational sectors and associated systems as discussed earlier, however, the study could not detect any effects on student achievement as a result of the policies influencing teaching and learning processes.

The arguments being presented suggest that although student achievement on the JLEE exams did show signs of improvement following the introduction of the NMC, this was due to adjustment of examination processes and was underscored by three key factors:

i. The three jumps that took place after the introduction of the NMC (2000 – 2010) compared to the single jump that took place before the policy introductions (Figure 9-4).
ii. The lack of improvement in aggregated achievement scores between 2001 and 2008.

iii. The uptick (0.62 ± 0.43 %p.a.) in the pass rate from one decade to the next (Figure 9-5).

The grade proportion distribution shows that the abrupt changes in the pass-fail rates for English, stated in (i) above, had a greater influence on the overall pass-fail rates, implying that each jump is reflective of systemic changes to the examination processes rather than an affective variation at the teaching and learning level.

One further point to be made was that the exams changed over the years in both construct and cognitive loads, and although the degree of variation determined from the analysis was not subtle, it did not represent large enough changes to explain the variations in achievement. Consequently, the analysis further supports arguments being made that background control mechanisms were being implemented that would explain the sudden jumps in achievement. When taken into consideration, as was the case between 2004 -2005, the trend showed general consistency in achievement rates (2001 -2008).

On the other hand, the outcome from this period does not support the idea that improvements in grade levels resulted from a lowering of quality or standards or the creation of easier examinations. Although there is an observable increase in the rate of improvement compared to the previous decade, this again was the result of the increased pass rates in 2001 and 2009 and associated with the systemic and abrupt changes in the mathematics and English exams.

Other factors working to improve the level of quality education as a result of the policies are seen in supporting documentation like “Knowing Our Schools” (MEYE, 2004b) which brought about added systemic branches within the education sector that enhanced quality education (in this case through quality control and auditing pressures).

10.6 Limitations

This study calls into question the effectiveness of broad-scale policies introduced to the Maltese educational system on student outcomes and consequently on student learning associated with the end of primary schooling. While multidimensional methods were used to support an objective understanding of impact, the results need to be interpreted with caution and recognise that their interpretations are limited.

The study relates specifically to the transitional stage between primary and secondary schooling and is established around a single set of examinations that were loosely considered benchmark exams. The JLEE exams, therefore, represent only part of the compulsory

schooling system implemented across the Maltese islands and thus reflect on that portion of the educational system alone. Other domains across the local educational landscape (early years, general primary, secondary attainment, post-secondary) would offer their own respective outcomes accordingly. Further research in these areas would expand the scope of these investigations to include other high-stakes examinations administered at different educational stages and should be established around systemic benchmarks. Benchmark assessments have been introduced to the Maltese system of education after 2010.

The study was also limited by a lack of formal interviews with policy developers, administrators, and educational directors from the period that the policies were introduced. The main reasons for this were time-related, with the policies having been introduced fifteen to twenty years before the study. Such methods would have strengthened an understanding of the interpretation of meaning and bolstered a broader means of analysing and explaining the results. Such discussions could have also guided the study towards other areas where records and information could be gathered regarding various other associated aspects —micro policies, memorandums, subject-specific actions, and adjustments.

Other limitations of the methodology were associated with the link between the psychometric data and the quality of education. Although the association can be made as argued in the literature section, it is done so within a framework of relative comparisons and not taken as a general measure of the quality of education at the time. The research could only render a relative judgment based on a longitudinal comparison for the same set of exams, once again, limiting the study to the JLEE bubble without being able to generalise or extrapolate to other areas in education.

From the outset, the study recognised that direct causal links between the policies and variation in outcomes could not be definitively established. Any direct association with actual student learning or other school improvements cannot be effectively made through this particular type of analyses (Feldhoff & Radisch, 2021). However, it is also understood that stronger threads of association need to be prepared during policy planning and development that would incorporate sets of indicators to apply as reflective monitoring tools.

10.7 Significance, contribution, and recommendations:

The literature and understanding of large-scale policy implementation recognise the complexities of such processes and argue consistently that these need to be accompanied by well-structured, data-driven monitoring systems. This research has highlighted that some

oversight was taking place, but none seemed to be concerned with measuring the impact on student learning.

This study goes some way towards responding to those shortcomings and although it gives clear signs of student progress following the introduction of the NMC and FACTS policies, it also indicates systemic issues that needed to be addressed to better inform the rollout of both these policies. These relate to the establishment of formal benchmark structures and broader educational effectiveness research. Other outcomes that can be drawn from this study consider the analytical validity of long-term longitudinal studies, the need for further research into CLT associated with assessment, and an understanding that readability studies for the Maltese language need further investigation and development.

The inputs informing this research were directly related to the two key policies. In that regard, the study suggests that effective cyclical monitoring and feedback systems would have played effective roles in delivering the main goals of those policies. As key information components, both these systems need to be integrated into future policies to support and enhance effective implementation. These would need to include the adoption of evidence-based policy development and monitoring systems (Slavin, 2020) directly associated with the impact on teaching and learning to guide the process. This argument stems mainly from the paucity of any records or literature assessing the impact of both the NMC and FACTS policies on student learning specifically. Although both policies recognised the importance of such systems (Mizzi, 1999), neither worked to integrate any such structures as an ongoing part of the implementation stage.

In considering the process and context issues associated with the JLEE, the main outcomes of the study indicate that the system lacked a fixed framework of examination standards against which to benchmark the examinations for purposes of monitoring and impact assessment. This hindered their ability to be used as an independent tool that could inform educational effectiveness in Malta. The findings also indicated that the JLEE was susceptible to possible alteration by the administering authorities, reducing the examination set's objectivity. This shortcoming seems to have been resolved with the NCF (MEDE, 2012), however, as the NCF fell outside the scope of this research, the study did not determine if those standards were implemented accordingly.

Turning next to arguments posited by Patrick, (1996) and Newton, (1997) that there is a time-dependent dilution of validity when comparing assessments longitudinally. Although the author agrees with the general principles of this premise, the study has raised questions about

how quickly that validity decays, especially if the context remains relatively consistent. Coe (2010) argued that performance-based comparisons can be applied if the exams had "comparable levels of challenge, complexity or skills". More relevant to this study, the data taken between 2001 and 2008 (spanning 7 years) indicate that under conditions of construct continuity and consistency, variations in outcome trends remained fairly steady. The relative uniformity of the constructs used by this study supports the argument that if construct validity remains continuous and test form complexity exhibits minimal variations, then any changes in student achievement can be considered a reflection of external influences including policy introduction. In the case of the JLEE, the introduced policies had little, if any, direct impact on student outcomes, meaning that there was no detectable influence on school learning environments that could have led to improved student performance on the assessment. The influence of the NMC and FACTS policies that actually led to an effect on student outcomes were determined to be related more to the systemic processes associated with the examination.

The fourth matter is associated with the multidimensional nature of the study that led to a deeper consideration of test form structures as they relate to cognitive loading on test takers. Research literature associated with cognitive load theory has mainly focused on teaching and learning environments with a smaller set of researchers having considered its significance for testing. The concept of consistent accessibility is however relatively important to test design when comparative longitudinal studies are being considered. The underpinning principles of CLT become relevant and exploitable for comparatively analysing sequences of test forms established on a common construct. This study needed to present arguments that the intrinsic and extraneous cognitive loads retained comparable levels over the period being studied and gross variation from the norm would have impacted comparative validity irrespective of the period.

The final issue being raised here suggests possible scope for further research into the use of LIX, ARI and CLI algorithms for determining the readability of Maltese texts. The study needed to develop its own algorithms to process non-English texts and although work has been done for other non-English texts little has been found relating to Maltese. The analysis using these three tools suggests that the application of language-independent algorithms delivered similar outcomes for the same texts and showed similar longitudinal trends for different texts over the fourteen years. The recommendation for further study would suggest that further work be considered to determine the validity of each of these algorithms but also introduce further

research into language-dependent algorithms for Maltese. Such research would allow better standardisation of comprehension texts and exam settings in the Maltese language.

The main recommendations can be summarised as follows:

1. Enhance the Monitoring Systems: The study emphasises the need for well-structured, data-driven monitoring systems to accompany large-scale policy implementation. These systems should measure the impact of policies on student learning and guide the implementation process. The study further suggests integrating effective cyclical monitoring and feedback systems into future policies.
2. Formal Benchmark Structures and Broader Educational Effectiveness Research: There needs to be an establishment of formal benchmark structures and broader educational effectiveness research to better inform policy rollouts.
3. Examination Standards: The study found that the JLEE lacked a fixed framework of examination standards for monitoring and impact assessment. A standardised examination framework should be developed and implemented to better inform an understanding of educational effectiveness.
4. Research on CLT and Maltese Text Readability: Enhance investigations into Cognitive Load Theory (CLT) in the context of assessments, and further develop readability studies for the Maltese language, including refining the LIX, ARI, and CLI algorithms used to determine readability.
5. Enhanced Understanding of Validity Decay in Longitudinal Studies: The study underscores the need for additional research to gain a deeper understanding of the rate at which comparative validity of examination outcomes diminishes in longitudinal studies, prompting questions about the pace of such validity decay.
6. Aggregated Grade Average Consideration: The study suggests considering an aggregated grade average rather than pass-fail rates as a benchmark indicator for student achievement.
7. Review Evaluation Methods: It may be beneficial to consider an aggregated grade average instead of only pass-fail rates as a benchmark indicator for tests. Similarly, exploring the use of a compensatory model for assessments could offer a more comprehensive perspective on student performance, potentially leading to a more nuanced understanding of achievement impact.

10.8 Concluding thoughts

The study set out to respond to a lack of research into the direct impact that nationwide, broad-scale policy introductions had on student achievement and attainment in Malta. It also worked to determine the practicality of analysing the JLEEs as an indicator of impact on student outcomes, how that could reflect on variations in learning, and subsequently, if they could be used to identify policy impact on the general quality of education. Originally, these ideas intended to include a broader scale analysis considering primary and secondary attainment, however, pragmatic challenges meant that the scope was reduced to focus on primary school attainment on the transitional examinations that were the JLEE. The importance of expanding the purpose of determining policy impact on achievement and attainment across all tiers of the educational system is still, however, considered to be crucial for continuously developing educational systems.

Although the study intended to establish the degree of success resulting from the reforms, the data was not sufficiently objective to establish a measure of effect. More specifically, although improvement could be detected and was, for example, shown to be approximately 0.2857 ± 0.1790 % p.a. between 2001 and 2008, this measure could not be claimed to be independent of institutional adjustments associated with the JLEE. Such systems need to be in place for the purpose of benchmarking or monitoring progress and must be allowed to function within a consistent framework independent of external modification. The evidence showing the influence of English and mathematics on pass-fail rates without there being direct knowledge of what varied in those examinations is one of the main points derived from this study. However, an objective assessment system based on established standards would allow and support mechanisms applied towards monitoring future change outcomes and identifying key aspects of what is driving that change.

The study also determined that although direct cause-and-effect links were not possible, the use of the i-p-o-c framework facilitated explanatory pathways that could be used to establish such connections. Coupled with impact assessment at the output level, such action should make it possible to support processes that would allow substantiated interpretation, informed decision making and managerial responsiveness. Similarly, it would support further investigative options and help determine whether the changes are having the desired effect at the micro and macro levels. Such a system would also support the possibility to analyse the impact on separate groups and categories of learners according to varying criteria and inform the decision-making process and responsiveness of the institutions as they implement future changes.

Further implications from this work suggest that the pass-fail rates alone cannot be used as a benchmark indicator for these tests. Rather, an aggregated grade average needs to be considered that does away with the otherwise skewed outcomes of the conjunctive criteria that affected student outcomes on the JLEE. The data has also shown that had a compensatory model been taken as the underpinning criteria for successfully passing the JLEE, then a higher annual proportion of students would have been successful over the years and variations in the success rates would have been more consistent. It is the opinion of this study that such assessment models established around examination sets would offer a truer picture of the impact on achievement if they were based on a compensatory aggregation of outcomes, rendering a more holistic understanding of effects taking place at the school level.

The results of this study support the idea that the influence of the introduced policies was not direct, rather, the policies themselves reflected an ongoing attempt by stakeholders at the time to implement paradigm shifts to various aspects of the educational landscape in Malta. This led to dynamic, evolving processes that were as much an output of the effort to change as the improvement in student achievement in the Junior Lyceum examinations. The policies tended to be applied principles that enable an evolving routine of implementation with different parts being applied and reapplied progressively as time went by and leading to a ubiquitous influence across the landscape to improve the system at large. It is the opinion of the author of this study that the policy influence was positive and could have had a stronger impact if defined targets had been established more explicitly.

11 References

Adams, D. (1993). Defining educational quality. *Improving Educational Quality Project Publication*, *1*. http://pdf.usaid.gov/pdf_docs/PNACA245.pdf

Adams Jr, J. E. (1994). Implementing program equity: Raising the stakes for educational policy and practice. *Educational Policy*, *8*(4), 518–534.

Aiken, L. R. (1979). Relationships between the item difficulty and discrimination indexes. *Educational and Psychological Measurement*, *39*(4), 821–824.

Aiken, W. M. (1942). The Story of the Eight-Year Study (Vol. 1). *New York*.

Alberts, R. V. (2001). Equating exams as a prerequisite for maintaining standards: Experience with Dutch centralised secondary examinations. *Assessment in Education: Principles, Policy & Practice*, *8*(3), 353–367.

American Educational Research Association , American Psychological Association, Joint Committee on Standards for Educational, Psychological Testing (US), & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. American Educational Research Association.

American Educational Research Association, The American Psychological Association, & The National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

American Institutes for Research. (1999). *Improving Educational Quality Project. Educational Quality Framework.* Report prepared for USAID.

Anderson, Bennett, N., & British Educational Management and Administration Society (Eds.). (2003). *Developing educational leadership: Using evidence for policy and practice*. Sage Publications.

Anderson, J. (1983). Lix and rix: Variations on a little-known readability index. *Journal of Reading*, *26*(6), 490–496.

Anderson, K., & Holloway, J. (2018). Discourse analysis as theory, method, and epistemology in studies of education policy. *Journal of Education Policy*, 1–34. https://doi.org/10.1080/02680939.2018.1552992

Anderson, L. W., & Bloom, B. S. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman,.

Anderson, R. C., & Davison, A. (1986). Conceptual and empirical bases of readability formulas. *Center for the Study of Reading Technical Report; No. 392*.

Anderson, V. R. (2016). Introduction to mixed methods approaches. *Handbook of Methodological Approaches to Community-Based Research: Qualitative, Quantitative, and Mixed Methods*, 233.

Angoff, W. H. (1984). *Scales, norms, and equivalent scores*. Educational Testing Service.

Archives New Zealand. (2007). *Digitisation Standard*. Archives New Zealand.

Armstrong, P. (2016). Bloom's taxonomy. *Vanderbilt University Center for Teaching*.

Astin, A. W., & Antonio, A. L. (2012). *Assessment for excellence: The philosophy and practice of assessment and evaluation in higher education* (2nd ed). Rowman & Littlefield Publishers.

Attard Tonna, M., & Bugeja, G. (2016). *A reflection on the learning outcomes framework project*.

Auld, E., Rappleye, J., & Morris, P. (2019). PISA for Development: How the OECD and World Bank shaped education governance post-2015. *Comparative Education*, *55*(2), 197–219.

Avis, J., Fisher, R., & Thompson, R. (2014). *Teaching in Lifelong Learning: A guide to theory and practice*. McGraw-Hill Education (UK).

Azorín, J. M., & Cameron, R. (2010). The application of mixed methods in organisational research: A literature review. *Electronic Journal of Business Research Methods*, *8*(2), 95–105.

Azzopardi, M., & Azzopardi, C. (2020). *Analysis of the discrimination index of final biology examinations in Malta*.

Baker, E. L. (2013). Critical Moments in Research and Use of Assessment. *Theory Into Practice*, *52*(sup1), 83–92. https://doi.org/10.1080/00405841.2013.795445

Ball, S. J. (1993). What is policy? Texts, trajectories and toolboxes. *The Australian Journal of Education Studies*, *13*(2), 10–17.

Bannert, M. (2002). Managing cognitive load—Recent trends in cognitive load theory. *Learning and Instruction*, *12*(1), 139–146.

Barber, M. (2010). How government, professions and citizens combine to drive successful educational change. In *Second international handbook of educational change* (pp. 261–278). Springer.

Bartholomew, T. T., & Brown, J. R. (2012). Mixed methods, culture, and psychology: A review of mixed methods in culture-specific psychological research. *International Perspectives in Psychology: Research, Practice, Consultation*, *1*(3), 177.

Bassey, M. (1999). *Case study research in educational settings*. McGraw-Hill Education (UK).

Beckmann, J. F., Birney, D. P., & Goode, N. (2017). Beyond psychometrics: The difference between difficult problem solving and complex problem solving. *Frontiers in Psychology*, *8*, 1739.

Beddow, P. A. (2018). Cognitive Load Theory for Test Design. In S. N. Elliott, R. J. Kettler, P. A. Beddow, & A. Kurz (Eds.), *Handbook of Accessible Instruction and Testing Practices: Issues, Innovations, and Applications* (pp. 199–211). Springer International Publishing. https://doi.org/10.1007/978-3-319-71126-3_13

Beddow, P. A., Kettler, R. J., & Elliott, S. N. (2008). Test accessibility and modification inventory (TAMI). *Peabody College, Vanderbilt University*.

Benavot, A. & UNESCO (Eds.). (2015). *Achievments and challenges* (1. ed). Unesco Publ.

Berman, P., & McLaughlin, M. W. (1974). *Federal Programs Supporting Educational Change: A Model of Educational Change. Volume I.*

Bezzina, C. (2003). The Maltese National Minimum Curriculum: The challenges ahead. *Management in Education*, *17*(1), 14–16.

Biglan, A., Ary, D., & Wagenaar, A. C. (2000). The value of interrupted time-series experiments for community intervention research. *Prevention Science*, *1*(1), 31–49.

Bird, E., Anderson, H. M., Anaya, G., & Moore, D. L. (2005). Beginning an assessment project: A case study using data audit and content analysis. *American Journal of Pharmaceutical Education*, *69*(1–5), 356.

Bodilly, S. (1996). *Lessons from New American Schools Development Corporation's Demonstration Phase.* ERIC.

Boeren, E. (2019). Understanding Sustainable Development Goal (SDG) 4 on "quality education" from micro, meso and macro perspectives. *International Review of Education*, *65*(2), 277–294. https://doi.org/10.1007/s11159-019-09772-7

Bogotch, I., Miron, L., & Biesta, G. (2007). "Effective for what; effective for whom?" Two questions SESI should not ignore. *International Handbook of School Effectiveness and Improvement*, 93–110.

Borg, C. (2004). *Message: Director General MEYE. In For All Children to Succeed. New Network Organisation For Quality Education In Malta*. Ministry of Education Youth and Employment.

Borg, M., & Giordmaina, J. (2012). *The College System In The State School Sector: A Study Of Its Impact As Perceived By College Principals, Members Of School Senior Management Teams, And Personnel In The Various Teaching Grades* (p. 328). MUT.

Borman, G. D., Hewes, G. M., Overman, L. T., & Brown, S. (2003). Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research*, *73*(2), 125–230.

Bowen, G. A. (2009). Document analysis as a qualitative research method. *Qualitative Research Journal*, *9*(2), 27–40.

Bowers, J. J. (1991). Evaluating testing programs at the state and local levels. *Theory into Practice*, *30*(1), 52–60.

Box-Steffensmeier, J. M., Freeman, J. R., Hitt, M. P., & Pevehouse, J. C. (2014). *Time series analysis for the social sciences*. Cambridge University Press.

Breakspear, S. (2012). *The Policy Impact of PISA* (OECD Education Working Papers No. 71). http://www.oecd-ilibrary.org/education/the-policy-impact-of-pisa_5k9fdfqffr28-en

Brindley, G. (1987). Factors affecting task difficulty. *Guidelines for the Development of Curriculum Resources*, 45–56.

Brucia, R. (2020). *Operationalizing item difficulty modeling in a medical certification context*. The University of North Carolina at Greensboro.

Bush, T. (2002). Educational Management: Theory and Practice. In *The principles and practice of educational management*. Sage.

Calleja, J., & Grima, G. (2012). *Message. In Ministry of Education and Employment, A National Curriculum Framework for All*. Ministry of Education and Employment.

Candlin, C. (1993). Task-based educational approaches. *Language Programs in Development Projects: Proceedings of the AIT RELC Conference .*, 225–237.

Caruana, C., & Allied Newspapers Ltd. (2016, July 15). *'Told you so', MUT says following poor exam results*. Times of Malta. https://www.timesofmalta.com/articles/view/20160715/local/told-you-so-mut-says-following-poor-exam-results.618889

Cassar, R. (2021). *Education and training policy*.

Cave, S. N., & von Stumm, S. (2021). Secondary data analysis of British population cohort studies: A practical guide for education researchers. *British Journal of Educational Psychology*, *91*(2), 531–546.

Chang, W.-C., & Chung, M.-S. (2009). Automatic applying Bloom's taxonomy to classify and analysis the cognition level of English question items. *2009 Joint Conferences on Pervasive Computing (JCPC)*, 727–734.

Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In 1. F. Bachman & AD Cohen (Eds.), Interfaces between second language acquisition and language testing (pp. 32-70). *New York: Cambridge University Pr*.

Chiavaroli, N., & Familari, M. (2011). When majority doesn't rule: The use of discrimination indices to improve the quality of MCQs. *Bioscience Education*, *17*(1), 1–7.

Clark, R. C., Nguyen, F., & Sweller, J. (2011). *Efficiency in learning: Evidence-based guidelines to manage cognitive load*. John Wiley & Sons.

Clune, W. H., White, P., & Patterson, J. H. (1989). *The implementation and effects of high school graduation requirements: First steps toward curricular reform*. Center for Policy Research in Education.

Codd, J. A. (1988). The construction and deconstruction of educational policy documents. *Journal of Education Policy*, *3*(3), 235–247.

Coe, R. (2010). Understanding comparability of examination standards. *Research Papers in Education*, *25*(3), 271–284.

Coe, R., Searle, J., Barmby, P., Jones, K., & Higgins, S. (2008). Relative difficulty of examinations in different subjects. *Durham: CEM Centre*.

Cohen, D. K., & Hill, H. C. (2001). *Learning policy: When state education reform works*. Yale University Press.

Coleman, M., & Lumby, J. (1999). The significance of site-based practitioner research in educational management. In *Practitioner research in education: Making a difference* (pp. 1–19).

Collins, L. M. (2006). Analysis of longitudinal data: The integration of theoretical model, temporal design, and statistical model. *Annu. Rev. Psychol.*, *57*, 505–528.

Commonwealth of Australia. (2013). *Digitising accumulated physical records*. National Archives of Australia. http://www.naa.gov.au/Images/dapr_tcm16-88758.docx

Cooper, A., Levin, B., & Campbell, C. (2009). The growing (but still limited) importance of evidence in education policy and practice. *Journal of Educational Change*, *10*(2–3), 159–171.

Crandall, D. P., & Loucks, S. F. (1983). *A Roadmap for School Improvement. Executive Summary of the Study of Dissemination Efforts Supporting School Improvement. People, Policies, and Practices: Examining the Chain of School Improvement, Volume X.*

Creemers, & Kyriakides. (2006). Critical analysis of the current approaches to modelling educational effectiveness: The importance of establishing a dynamic model. *School Effectiveness and School Improvement*, *17*(3), 347–366. https://doi.org/10.1080/09243450600697242

Creemers, & Kyriakides. (2007). *The dynamics of educational effectiveness: A contribution to policy, practice and theory in contemporary schools*. Routledge.

Crisp, V., & Novaković, N. [zbreve] da. (2009). Is this year's exam as demanding as last year's? Using a pilot method to evaluate the consistency of examination demands over time. *Evaluation & Research in Education*, *22*(1), 3–15.

Cristina, D. (2012). *Message. In Ministry of Education and Employment, A National Curriculum Framework for All*. Ministry of Education and Employment.

Curriculum Department, & Educational Assessment Unit. (1997). *Entrance Examinations into Junior Lyceums*. Ministry of Education.

Curriculum Department, & Educational Assessment Unit. (1998). *Entrance Examinations into Junior Lyceums*. Ministry of Education.

Curriculum Department, & Educational Assessment Unit. (1999). *Entrance Examinations into Junior Lyceums*. Ministry of Education.

Curriculum Department, & Educational Assessment Unit. (2000). *Entrance Examinations into Junior Lyceums*. Ministry of Education.

Curriculum Department, & Educational Assessment Unit. (2002). *Entrance Examinations into Junior Lyceums*. Ministry of Education.

Curriculum Department, & Educational Assessment Unit. (2004). *Entrance Examinations into Junior Lyceums*. Ministry of Education.

Curriculum Department, & Educational Assessment Unit. (2005). *Entrance Examinations into Junior Lyceums*. Ministry of Education.

Curriculum Department, & Educational Assessment Unit. (2006). *Entrance Examinations into Junior Lyceums*. Ministry of Education.

Curriculum Department, & Educational Assessment Unit. (2007). *Entrance Examinations into Junior Lyceums*. Ministry of Education.

Curriculum Department, & Educational Assessment Unit. (2008). *Entrance Examinations into Junior Lyceums*. Ministry of Education.

Curriculum Department, & Educational Assessment Unit. (2009). *Entrance Examinations into Junior Lyceums*. Ministry of Education.

Curriculum Department, & Educational Assessment Unit. (2010). *Entrance Examinations into Junior Lyceums*. Ministry of Education.

Cutajar, M. (2007). Educational reform in the Maltese Islands. *Journal of Maltese Education Research*, *5*(1), 3–21.

Darling-Hammond, L. (2010). Teaching and educational transformation. In *Second international handbook of educational change* (pp. 505–520). Springer.

Datnow, A. (2006). Connections in the policy chain. *New Directions in Education Policy Implementation Confronting Complexity*, 105–124.

Datnow, A., & Park, V. (2010). Large-scale reform in the era of accountability: The system role in supporting data-driven decision making. In *Second international handbook of educational change* (pp. 209–220). Springer.

DeLeeuw, K. E., & Mayer, R. E. (2008). A comparison of three measures of cognitive load: Evidence for separable measures of intrinsic, extraneous, and germane load. *Journal of Educational Psychology*, *100*(1), 223.

Desforges, C. (2003). Evidence -Informed Policy and Practice in Teaching and Learning. In L. Anderson, N. Bennett, & British Educational Management and Administration Society (Eds.), *Developing educational leadership: Using evidence for policy and practice* (pp. 3–10). Sage Publications.

DiBattista, D., & Kurzawa, L. (2011). Examination of the quality of multiple-choice items on classroom tests. *Canadian Journal for the Scholarship of Teaching and Learning*, *2*(2), 4.

Doneva, R., Gaftandzhieva, S., & Totkov, G. (2018). Automated Quality Assurance of Educational Testing. *Turkish Online Journal of Distance Education*, *19*(3), 71–92.

Dorans, N. J., Moses, T. P., & Eignor, D. R. (2010). Principles and practices of test score equating. *ETS Research Report Series*, *2010*(2), i–41.

Dorans, N. J., Pommerich, M., & Holland, P. W. (2007). *Linking and aligning scores and scales*. Springer Science & Business Media.

Doublesin, G. (2022). *Visualising trends in examination standards: Patterns in Discrimination vs Facility plots*. https://doi.org/10.35542/osf.io/s3p75

Douglas, K. M., & Mislevy, R. J. (2010). Estimating classification accuracy for complex decision rules based on multiple scores. *Journal of Educational and Behavioral Statistics*, *35*(3), 280–306.

Downing, S. M. (2003). Validity: On the meaningful interpretation of assessment data. *Medical Education*, *37*(9), 830–837.

DQSE : Directorate for Quality and Standards in Education. (2016).

Dueñas, G., Jimenez, S., & Baquero, J. (2015). Automatic prediction of item difficulty for short-answer questions. *2015 10th Computing Colombian Conference (10CCC)*, 478–485.

Duffy, F. M., Reigeluth, C. M., Solomon, M., Caine, G., Carr-Chellman, A. A., Almeida, L., Frick, T., Thompson, K., Koh, J., & Ryan, C. D. (2006). The process of systemic change. *TechTrends*, *50*(2), 41–51.

EAU : Educational Assessment Unit. (n.d.).

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5 ed). Prentice-Hall Englewood Cliffs, NJ.

Education Statistics 2006. (2010). National Statistics Office. https://nso.gov.mt/wp-content/uploads/education06_08.pdf

Educational Assessment Unit. (1997–2010). *Junior Lyceum Entrance Exam: Results Register*. Ministry of Education.

Educational Assessment Unit. (n.d.). *GUIDELINES FOR PAPER SETTERS*. https://curriculum.gov.mt/en/Assessment/Assessment-of-Learning/Documents/guidelines_paper_setters.pdf

Embretson, S. E., & Wetzel, C. D. (1987). Component latent trait models for paragraph comprehension tests. *Applied Psychological Measurement*, *11*(2), 175–193.

Escudero, E. B., Reyna, N. L., & Morales, M. R. (2000). The level of difficulty and discrimination power of the Basic Knowledge and Skills Examination (EXHCOBA). *Revista Electrónica de Investigación Educativa*, *2*(1), 2.

Ethical Guidelines for Educational Research, fourth edition. (2018). https://www.bera.ac.uk/researchers-resources/publications/ethical-guidelines-for-educational-research-2018

European Commission. Education, A. and C. E. A. (EACEA). (2017). *Support mechanisms for evidence-based policy-making in education*.

European Commission/EACEA/Eurydice. (2017). Support Mechanisms for Evidence-Based Policy-Making in Education. Eurydice Report. *Education, Audiovisual and Culture Executive Agency, European Commission*.

Feldhoff, T., & Radisch, F. (2021). Why must everything be so complicated? Demands and challenges on methods for analyzing school improvement processes. *Concept and Design Developments in School Improvement Research*, 9.

Fenech Adami, A. (2004). Enhancing students' learning through differentiated approaches to teaching and learning: A Maltese perspective. *Journal of Research in Special Educational Needs*, *4*(2), 91–97. https://doi.org/10.1111/j.1471-3802.2004.00023.x

Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. F., & Hemphill, C. F. (1998). *Uncommon measures: Equivalence and linkage among educational tests*. National Academies Press.

Fiske, E., B., & UNESCO. (2000). *World Education Forum, Dakar, Senegal, 26-28 April 2000: Final report*. UNESCO.

Fitz-Gibbon, C. T., & Vincent, L. (1997). Difficulties regarding subject difficulties: Developing reasonable explanations for observable data. *Oxford Review of Education*, *23*(3), 291–298.

Frick, T., Thompson, K., & Koh, J. (2015). Systemic Change: Get Ready, SET, Go!–Where? *TechTrends*, *50*(2), 47–48.

Fuhrman, S., Clune, W., & Elmore, R. (1988). Research on education reform: Lessons on the implementation of policy. *Teachers College Record*, *90*(2), 237–257.

Fullan, M. (1993). *Change forces: Probing the depths of educational reform* (Vol. 10). Psychology Press.

Galea, L. (1999). *Forward. In Ministry of Education Youth and Employment, Creating the Future Together. National Minimum Curriculum*. Ministry of Education Youth and Employment.

Galea, L. (2004). *Forward. In Ministry of Education Youth and Employment, For All Children to Succeed*. Ministry of Education Youth and Employment.

Gall, M. D., Borg, W. R., & Gall, J. P. (1996). *Educational research: An introduction*. Longman Publishing.

Gerberich, J. R. (1963). The development of educational testing. *Theory into Practice*, *2*(4), 184–191.

Gillmor, S. C., Poggio, J., & Embretson, S. (2015). Effects of reducing the cognitive load of mathematics test items on student performance. *Numeracy*, *8*(1), 4.

Ginsburg, M. B., Cooper, S., Raghu, R., & Zegarra, H. (1990). National and world-system explanations of educational reform. *Comparative Education Review*, *34*(4), 474–499.

Glass, G. V. (1997). Interrupted time series quasi-experiments. *Complementary Methods for Research in Education*, *2*, 589–608.

Glass, G. V. (2006). Interrupted Time Series Quasi-Experiments. In J. L. Green, G. Camilli, P. B. Elmore, & American Educational Research Association (Eds.), *Handbook of complementary methods in education research* (pp. 589–608). Lawrence Erlbaum Associates ; Published for the American Educational Research Association.

Goldstein, H. (2001). Using pupil performance data for judging schools and teachers: Scope and limitations. *British Educational Research Journal*, *27*(4), 433–442.

Goldstein, H., & Cresswell, M. (1996). The comparability of different subjects in public examinations: A theoretical and practical critique. *Oxford Review of Education*, *22*(4), 435–442.

Gorard, S. (2001). *A changing climate for educational research? The role of research capability-building*.

Gorard, S., See, B. H., & Siddiqui, N. (2017). *The trials of evidence-based education*. Routledge.

Gray, J., Goldstein, H., & Thomas, S. (2003). Of trends and trajectories: Searching for patterns in school improvement. *British Educational Research Journal*, *29*(1), 83–88.

Grima, G., Grech, L., Mallia, C., Mizzi, B., Vassallo, P., & Ventura, F. (2008). *Transition from primary to secondary schools in Malta: A review*. Floriana, Malta: Ministry for Education, Youth and Sport.

Hakim, C. (1982). Secondary analysis and the relationship between official and academic social research. *Sociology*, *16*(1), 12–28.

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. Routledge.

Halpin, D. (1994). Practice and prospects in education policy research. *Researching Education Policy: Ethical and Methodological Issues*, *14*, 205.

Hamann, E. T., & Lane, B. (2004). The roles of state departments of education as policy intermediaries: Two cases. *Educational Policy*, *18*(3), 426–455.

Hamersley, M. (2008). Causality as Conundrum: The Case of Qualitative Inquiry. *Methodological Innovations Online*, *2*(3), 1–5. https://doi.org/10.4256/mio.2008.0001

Hancock, G. R. (1994). Cognitive complexity and the comparability of multiple-choice and constructed-response test formats. *The Journal of Experimental Education*, *62*(2), 143–157.

Hanson, B. (1993). Equipercentile equating with equal interval scores. *Unpublished Manuscript*.

Hanushek, E. A., & Wößmann, L. (2007). *The role of education quality for economic growth*. The World Bank.

Hargreaves, A., Lieberman, A., Fullan, M., & Hopkins, D. W. (2014). *International handbook of educational change: Part two* (Vol. 5). Springer.

Harvey, L., & Green, D. (1993). Defining Quality. *Assessment & Evaluation in Higher Education*, *18*(1), 9–34. https://doi.org/10.1080/0260293930180102

Healey, F. H., & DeStefano, J. (1997). *Education Reform Support: A Framework for Scaling Up School Reform. Policy Paper Series.*

Henry, M. (1993). What is policy? A response to Stephen Ball. *Discourse*, *14*(1), 102–105.

Herscovics, N., & Linchevski, L. (1994). A cognitive gap between arithmetic and algebra. *Educational Studies in Mathematics*, *27*(1), 59–78.

Hewitt, M. A., & Homan, S. P. (2003). Readability level of standardized test items and student performance: The forgotten validity variable. *Literacy Research and Instruction*, *43*(2), 1–16.

Higgins, S. (2020). The development and worldwide impact of the Teaching and Learning Toolkit. *Getting Evidence into Education. Evaluating the Routes to Policy & Practice*, 69–83.

Hill, M., & Varone, F. (2016). *The public policy process*. Routledge.

Hitch, G. J. (1978). The role of short-term working memory in mental arithmetic. *Cognitive Psychology*, *10*(3), 302–323.

Honig. (2009). What works in defining "what works" in educational improvement: Lessons from educational policy implementation research. Directions for future research. *Handbook of Educational Policy Research. New York: Routledge*, 333–347.

Honig, & Coburn. (2008a). Evidence-based decision making in school district central offices: Toward a policy and research agenda. *Educational Policy*, *22*(4), 578–608.

Honig, M., I., & Coburn, C. (2008b). Evidence-based decision making in school district central offices: Toward a policy and research agenda. *Educational Policy*, *22*(4), 578–608.

Hopkins, D., Stringfield, S., Harris, A., Stoll, L., & Mackay, T. (2014). School and system improvement: A narrative state-of-the-art review. *School Effectiveness and School Improvement*, *25*(2), 257–281.

Hotiu, A. (2006). *The relationship between item difficulty and discrimination indices in multiple-choice tests in a physical science course*. Citeseer.

Irwing, P., & Hughes, D. J. (2018). Test Development. *The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development*, 1–47.

Iwu, C. G., Ezeuduji, I. O., Iwu, I. C., Ikebuaku, K., & Tengeh, R. K. (2018). Achieving quality education by understanding teacher job satisfaction determinants. *Social Sciences*, *7*(2), 25.

Jandaghi, G., & Shaterian, F. (2008). Validity, Reliability and Difficulty Indices for Instructor-Built Exam Questions. *Journal of Applied Quantitative Methods*, *3*(2), 151–155.

Jones, B., & Ratcliffe, P. (1996). Comparing the Standards of Examining Groups in the United Kingdom. *Issues in Setting Standards: Establishing Comparabilities*, 57.

Jones, Harland, J., Reid, J. M., & Bartlett, R. (2009). Relationship between examination questions and bloom's taxonomy. *2009 39th IEEE Frontiers in Education Conference*, 1–6.

Joseph, R., & Reigeluth, C. M. (2010). The systemic change process in education: A conceptual framework. *Contemporary Educational Technology*, *1*(2).

Joshi, P. K., Jain, Y., Khunyakari, R., & Basu, S. (2020). An Alternative Approach to Calculate Discrimination Index. *GPG Journal Of*.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73.

Karelia, B. N., Pillai, A., & Vegada, B. N. (2013). The levels of difficulty and discrimination indices and relationship between them in four-response type multiple choice questions of pharmacology summative tests of year II MBBS students. *IeJSME*, *7*(2), 41–46.

Kettler, R. J., Elliott, S. N., & Beddow, P. A. (2009). Modifying achievement test items: A theory-guided and data-based approach for better measurement of what students with disabilities know. *Peabody Journal of Education*, *84*(4), 529–551.

Kettler, R. J., Elliott, S. N., Beddow, P. A., & Kurz, A. (2018). Accessible instruction and testing today. In *Handbook of Accessible Instruction and Testing Practices* (pp. 1–16). Springer.

Khoshaim, H. B., & Rashid, S. (2016). Assessment of the Assessment Tool: Analysis of Items in a Non-MCQ Mathematics Exam. *International Journal of Instruction*, *9*(1), 119–132.

Kickert, R., Meeuwisse, M., Arends, L. R., Prinzie, P., & Stegers-Jager, K. M. (2021). Assessment policies and academic progress: Differences in performance and selection for progress. *Assessment & Evaluation in Higher Education*, *46*(7), 1140–1156.

King, K. (2007). Multilateral agencies in the construction of the global agenda on education. *Comparative Education*, *43*(3), 377–391.

Knuth, E. J., Alibali, M. W., McNeil, N. M., Weinberg, A., & Stephens, A. C. (2005). Middle school students' understanding of core algebraic concepts: Equivalence & Variable1. *Zentralblatt Für Didaktik Der Mathematik*, *37*(1), 68–76.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling: Methods and practices*. New York, NY: Springer.

Kontopantelis, E., Doran, T., Springate, D. A., Buchan, I., & Reeves, D. (2015). Regression based quasi-experimental approach when randomisation is not an option: Interrupted time series analysis. *Bmj*, *350*, h2750.

Kreber, C., & Brook, P. (2010). Impact evaluation of educational development programmes. *International Journal for Academic Development*, *6*(2), 96–108. https://doi.org/10.1080/13601440110090749

Krell, M. (2017). Evaluating an instrument to measure mental load and mental effort considering different sources of validity evidence. *Cogent Education*, *4*(1), 1280256.

Krippendorff, K. (1989). Content analysis. In E. Barnouw, G. Gerbner, W. Schramm, T. L. Worth, & L. Gross (Ed.), *Content Analaysis* (Vol. 1, pp. 403–407). http://repository.upenn.edu/asc_papers/226

Lagarde, M. (2012). How to do (or not to do)… Assessing the impact of a policy change with routine longitudinal data. *Health Policy and Planning*, *27*(1), 76–83.

Laird, N. M., & Mosteller, F. (1990). Some statistical methods for combining experimental results. *International Journal of Technology Assessment in Health Care*, *6*(1), 5–30.

Lee, F.-L., & Heyworth, R. (2000). *Problem complexity: A measure of problem difficulty in algebra by using computer*.

Lennon, M. C. (2016). *In search of quality: Evaluating the impact of learning outcomes policies in higher education regulation*. University of Toronto (Canada).

Levenson, M. S., Banks, D. L., Eberhardt, K. R., Gill, L. M., Guthrie, W. F., Liu, H., Vangel, M. G., Yen, J. H., & Zhang, N. F. (2000). An approach to combining results from multiple methods motivated by the ISO GUM. *Journal of Research of the National Institute of Standards and Technology*, *105*(4), 571.

Levin, B. (2010). How to change 5,000 schools. In *Second international handbook of educational change* (pp. 309–322). Springer.

Lewis, S., & Hogan, A. (2019). Reform first and ask questions later? The implications of (fast) schooling policy and 'silver bullet'solutions. *Critical Studies in Education*, *60*(1), 1–18.

Liu, J., & Dorans, N. J. (2016). Fairness in score interpretation. *Fairness in Educational Assessment and Measurement*, 77–96.

Livingston, S. A. (2004). *Equating test scores (without IRT*.

Logan, T. (2020). A practical, iterative framework for secondary data analysis in educational research. *The Australian Educational Researcher*, *47*(1), 129–148.

Louis, K. S. (2010). Better schools through better knowledge? New understanding, new uncertainty. In *Second international handbook of educational change* (pp. 3–27). Springer.

Luo, S., & Skehan, P. (2007). *Re-examining factors that affect task difficulty in TBLA*. Chinese University of Hong Kong.

MacBeath, J. (2007). Improving school effectiveness: Retrospective and prospective. *International Handbook of School Effectiveness and Improvement*, 57–74.

Malen, B. (2006). Revisiting policy implementation as a political phenomenon. *New Directions in Education Policy Implementation*, 83–104.

Marston, D. (1988). The effectiveness of special education: A time series analysis of reading performance in regular and special education settings. *The Journal of Special Education*, *21*(4), 13–26.

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist*, *34*(4), 207–218.

Matlock-Hetzel, S. (1997). *Basic Concepts in Item and Test Analysis.*

McBee, M. T., Peters, S. J., & Waterman, C. (2014). Combining scores in multiple-criteria assessment systems: The impact of combination rule. *Gifted Child Quarterly*, *58*(1), 69–89.

McCowan, R. J., & McCowan, S. C. (1999). Item Analysis for Criterion-Referenced Tests. *Online Submission*.

McLaughlin, M. W. (1987). Learning from experience: Lessons from policy implementation. *Educational Evaluation and Policy Analysis*, *9*(2), 171–178.

MEDE. (2012). *A national curriculum framework for all* (December 2012).

Melovitz Vasan, C. A., DeFouw, D. O., Holland, B. K., & Vasan, N. S. (2018). Analysis of testing with multiple choice versus open-ended questions: Outcome-based observations in an anatomy course. *Anatomical Sciences Education*, *11*(3), 254–261.

Metsämuuronen, J. (2018). Generalized Discrimination Index. *DOI: Http://Dx. Doi. Org/10.13140/RG*, *2*(30933.88804), 1.

MEYE. (1999). *Creating the Future Together. National Minimum Curriculum*. Ministry of Education Floriana,, Malta.

MEYE. (2004a). *For All Children to Succeed, A New Network Organisation for Quality Education in Malta.* Ministry of Education, Youth and Employment.

MEYE. (2004b). *Knowing Our Schools*. Department of Operations Education Division, Ministry of Education, Youth and Employment, Malta.

Mifsud, D. (2015). Policy-mandated collegiality in the Maltese education scenario: The experience of the leaders. *Malta Review of Educational Research*, *9*(2), 209–226.

Mifsud, G. (2019). *The readability of Maltese examination texts*.

Miller, C. (2011). Aesthetics and e-assessment: The interplay of emotional design and learner performance. *Distance Education*, *32*(3), 307–337.

Mizzi, C. (1999). *Message: Director General Of Education. In Creating the Future Together. National Minimum Curriculum*. Ministry of Education Youth and Employment.

Mizzi, C. (2004). *Message: Permanent Secretary MEYE. In For All Children to Succeed. New Network Organisation For Quality Education In Malta*. Ministry of Education Youth and Employment.

Moeini, R. (2020). *Cognitive evidence for construct validity of the IELTS Reading Comprehension Module: Content analysis, test taking processes, and experts' accounts*. Carleton University.

Morrison, M. (2007). What do we mean by educational research. *Research Methods in Educational Leadership and Management*, *2*, 13–36.

Mullis, I. V., & Jenkins, L. B. (1990). *The Reading Report Card, 1971-88: Trends from the Nation's Report Card.* ERIC.

Musa, A., Shaheen, S., Elmardi, A., & Ahmed, A. (2018). Item difficulty & item discrimination as quality indicators of physiology MCQ examinations at the Faculty of Medicine Khartoum University. *Khartoum Medical Journal*, *11*(2).

My Byline Media. (2020). *Check your Readability.* [Readability Formulas:]. AUTOMATIC READABILITY CHECKER. https://www.readabilityformulas.com/free-readability-formula-tests.php

Newman, D. L., Kundert, D. K., Lane Jr, D. S., & Bull, K. S. (1988). Effect of varying item order on multiple-choice test scores: Importance of statistical and cognitive difficulty. *Applied Measurement in Education*, *1*(1), 89–97.

Newton, P. (1997). Examining standards over time. *Research Papers in Education*, *12*(3), 227–247.

Newton, P. (2005). Examination standards and the limits of linking. *Assessment in Education: Principles, Policy & Practice*, *12*(2), 105–123.

Newton, P. (2021). Demythologising A level Exam Standards. *Research Papers in Education*, 1–32.

Nunan, D., & Keobke, K. (1995). Task Difficulty from the Learner's Perspective: Perceptions and Reality. *Hong Kong Papers in Linguistics and Language Teaching*, *18*, 1–12.

Nunnery, J. A. (1998). Reform Ideology and the Locus of Development Problem in Educational Restructuring: Enduring Lessons from Studies of Educational Innovation. *Education and Urban Society*, *30*(3), 277–295.

OECD (Ed.). (2013). *Synergies for better learning: An international perspective on evaluation and assessment*. OECD.

OECD. (2018). *Education at a Glance 2018*. https://www.oecd-ilibrary.org/content/publication/eag-2018-en

OECD Review on Evaluation and Assessment Frameworks for Improving School Outcomes—OECD. (n.d.). Retrieved 2 April 2017, from

http://www.oecd.org/edu/school/oecdreviewonevaluationandassessmentframeworksforimprovingschooloutcomes.htm

OECD, Ÿ. (2012). *Equity and quality in education: Supporting disadvantaged students and schools*. OECD Publishing Paris.

Operations and Programme Implementation Directorate (OPM). (2007). *Annual Reports of Government Departments—2006*.

Paas, F. G., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist*, *38*(1), 63–71.

Paas, F. G., & Van Merriënboer, J. J. (1994). Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Review*, *6*(4), 351–371.

Panchenko, L., & Samovilova, N. (2020). Secondary data analysis in educational research: Opportunities for PhD students. *SHS Web of Conferences*, *75*, 04005.

Patrick, H. (1996, September). Comparing Public Examination Standards over Time. *Paper Presented at BERA Conference, (Sept1996)*. BERA Conference,.

Pelánek, R., Effenberger, T., & Čechák, J. (2022). Complexity and difficulty of items in learning systems. *International Journal of Artificial Intelligence in Education*, *32*(1), 196–232.

Pellegrini, M., & Vivanet, G. (2021). Evidence-based policies in education: Initiatives and challenges in Europe. *ECNU Review of Education*, *4*(1), 25–45.

Pisani, M., Cassar, C. M., & Muscat, V. (2010). A review of the national minimum curriculum from an equality perspective. *Malta: National Commission for the Promotion of Equality*.

Policy Documentation Archive. (n.d.). Retrieved 22 September 2022, from https://education.gov.mt/en/resources/Pages/Policy-Documentation-Archive.aspx

Pommerich, M. (2016). The fairness of comparing test scores across different tests or modes of administration. *Fairness in Educational Assessment and Measurement*, 111–134.

Popham, W. J. (1999). Why standardized tests don't measure educational quality. *Educational Leadership*, *56*, 8–16.

Purpura, J. E. (2013). Assessing grammar. *The Companion to Language Assessment*, *1*, 100–124.

Pyrczak, F. (1973). VALIDITY OF THE DISCRIMINATION INDEX AS A MEASURE OF ITEM QUALITY 1. *Journal of Educational Measurement*, *10*(3), 227–231.

Reck, R. P., & Reck, R. A. (2007). Generating and rendering readability scores for Project Gutenberg texts. *Proceedings of the Corpus Linguistics Conference*.

Reigeluth, C. M. (2006a). A leveraged emergent approach to systemic transformation. *Tech Trends*, *50*(2), 46–47.

Reigeluth, C. M. (2006b). The Guidance system for transforming education. *Tech Trends*, *50*(2), 42.

Rudolph, M. J., Daugherty, K. K., Ray, M. E., Shuford, V. P., Lebovitz, L., & DiVall, M. V. (2019). Best practices related to examination item construction and post-hoc review. *American Journal of Pharmaceutical Education*, *83*(7).

Rush, B. R., Rankin, D. C., & White, B. J. (2016). The impact of item-writing flaws and item complexity on examination item difficulty and discrimination value. *BMC Medical Education*, *16*(1), 250.

Sackney, L. (2007). History of the school effectiveness and improvement movement in Canada over the past 25 years. *International Handbook of School Effectiveness and Improvement*, 167–182.

Sahlberg, P. (2010). Educational change in Finland. In *Second international handbook of educational change* (pp. 323–348). Springer.

Sahlberg, P. (2017). *The global testing culture: Shaping education policy, perceptions, and practice edited by William C. Smith*. Taylor & Francis.

Sahney, S., Banwet, D. K., & Karunes, S. (2008). An integrated framework of indices for quality management in education: A faculty perspective. *The TQM Journal*, *20*(5), 502–519. https://doi.org/10.1108/17542730810898467

Said, L. (2015). The Influence of Teacher Behaviours on Pupils' Mathematical Attainment at Age 6-Said.pdf. *Malta Review of Educational Research*, *9*(2), 331–353.

Salkind, N. J. (2010). *Encyclopedia of research design* (Vol. 1). Sage.

Sammons, P. (2009). The dynamics of educational effectiveness: A contribution to policy, practice and theory in contemporary schools. *School Effectiveness and School Improvement*, *20*(1), 123–129. https://doi.org/10.1080/09243450802664321

Santiago, P. (n.d.). *Evaluation and Assessment Frameworks for Improving School Outcomes—Common Policy Challenges*. OECD.

Sax, G., Eilenberg, E. G., & Klockars, A. J. (1972). Achievement as a function of test item complexity and difficulty. *The Journal of Experimental Education*, *40*(4), 90–93.

Scheerens, J. (2004). The conceptual framework for measuring quality. *Recuperado de: Http://Lascuolachefunziona. Pbworks. Com/f/An% 20input-Processoutcome% 20framework% 20for% 20assessing% 20education% 20quality. Pdf*.

Scheerens, J., Luyten, H., & van Ravens, J. (2011a). Measuring Educational Quality by Means of Indicators. In J. Scheerens, H. Luyten, & J. van Ravens (Eds.), *Perspectives on Educational Quality* (Vol. 1, pp. 35–50). Springer Netherlands. http://link.springer.com/10.1007/978-94-007-0926-3_2

Scheerens, J., Luyten, H., & van Ravens, J. (2011b). *Perspectives on educational quality: Illustrative outcomes on primary and secondary schooling in the Netherlands*. Springer.

Scott, J. (1990). *A matter of record*. Cambridge: Polity Press.

Sebba, J. (2003). A Government Strategy for Research and Development in Education. In L. Anderson, N. Bennett, & British Educational Management and Administration Society (Eds.), *Developing educational leadership: Using evidence for policy and practice* (pp. 11–24). Sage Publications.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Wadsworth Cengage learning.

Shadish, W. R., Leullen, J. K., Green, J. L., Camilli, G., & Elmore, P. B. (2006). Quasi-Experimental Design. In *Handbook of complementary methods in education research*.

Shaw, S., & Crisp, V. (2012). An approach to validation: Developing and applying an approach for the validation of general qualifications. *Research Matters, Special*, *3*, 1–44.

Siddiqui, N. (2019). *Using secondary data in education research • Understanding the forms of secondary datasets • Strengths and limitations of secondary data resources • Linking secondary data for research*.

Sim, S.-M., & Rasiah, R. I. (2006). Relationship between item difficulty and discrimination indices in true/false-type multiple choice questions of a para-clinical multidisciplinary paper. *Annals-Academy of Medicine Singapore*, *35*(2), 67.

Simkin, M. G., & Kuechler, W. L. (2005). Multiple-choice tests and student understanding: What is the connection? *Decision Sciences Journal of Innovative Education*, *3*(1), 73–98.

Singer, J. D., Willett, J. B., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford university press.

Slavin, R. E. (2020). How evidence-based reform will transform research and practice in education. *Educational Psychologist*, *55*(1), 21–31.

Smith, E. (2008). Pitfalls and promises: The use of secondary data analysis in educational research. *British Journal of Educational Studies*, *56*(3), 323–339.

Smith, M. S., & O'Day, J. (1990). *Systemic school reform*.

Split Text—Online Text Tools. (n.d.). Retrieved 31 December 2020, from https://onlinetexttools.com/split-text

Squire, K. D., & Reigeluth, C. M. (2000). The many faces of systemic change. *Educational Horizons*, *78*(3), 143–152.

Stanny, C. J. (2016). Reevaluating Bloom's Taxonomy: What measurable verbs can and cannot say about student learning. *Education Sciences*, *6*(4), 37.

Stephens, D. (2003). Quality of basic education. *Paper for EFA Global Monitoring Report, Http://Unesdoc. Unesco. Org/Images/0014/001469/146968e. Pdf*.

Stevenson, H. (2003). *Educational Policy,*. CELM.

Stone, D. (2001). Getting research into policy. *Third Annual Global Development Network Conference on 'Blending Local and Global Knowledge', Rio De Janeiro*, *10*.

Stringfield, S. (1997). *Urban and suburban/rural special strategies for educating disadvantaged children: Second year report*. Planning and Evaluation Service, US Dept. of Education.

Supovitz, J. A., & Taylor, B. S. (2005). Systemic Education Evaluation Evaluating the Impact of Systemwide Reform in Education. *American Journal of Evaluation*, *26*(2), 204–230.

Suruchi, S., & Rana, S. S. (2014). Test item analysis and relationship between difficulty level and discrimination index of test items in an achievement test in biology. *PIJR*, *3*(6), 56–58.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, *4*(4), 295–312.

Sweller, J., Van Merriënboer, J. J. G., & Paas, F. G. (1998). Cognitive Architecture and Instructional Design. *Educational Psychology Review*, *10*(3), 251–296.

Sweller, J., van Merriënboer, J. J., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, *31*(2), 261–292.

Taylor, S., Lingard, B., Rizvi, F., & Henry, M. (1997a). Doing policy Analysis. In *Educational policy and the politics of change* (pp. 36–53). Routledge.

Taylor, S., Lingard, B., Rizvi, F., & Henry, M. (1997b). *Educational policy and the politics of change*. Routledge.

Teddlie, C., & Stringfield, S. (2007). A history of school effectiveness and improvement research in the USA focusing on the past quarter century. In *International handbook of school effectiveness and improvement* (pp. 131–166). Springer.

Teddlie, C., & Tashakkori, A. (2009). *Foundations of mixed methods research: Integrating quantitative and qualitative approaches in the social and behavioral sciences*. Sage.

The National Archives. (2016). *Digitisation at The National Archives*. http://www.nationalarchives.gov.uk/documents/information-management/digitisation-at-the-national-archives.pdf

Thomas, R. M. (2002). *Overcoming inertia in school reform: How to successfully implement change*. Corwin Press.

Tillman, R., & Hagberg, L. (2014). *Readability algorithms compability on multiple languages*.

Townsend, T. (2007). 20 years of ICSEI: The impact of school effectiveness and school improvement on school reform. In *International handbook of school effectiveness and improvement* (pp. 3–26). Springer.

UNESCO. (1990). *World declaration on education for all and framework for action to meet basic learning needs.* Inter-Agency Commission. //catalog.hathitrust.org/Record/010577427

UNESCO. (2000). *Dakar Framework of Action*.

UNESCO. (2002). *Education for all: Is the world on track?.* UNESCO.

UNESCO. (2005). *The quality imperative: Education for All* (2nd printing).

UNESCO. (2015). *Education 2030 Incheon Declaration and Framework for Action*. UNESCO Paris.

UNESCO. (2017). *Education for sustainable development goals: Learning objectives*. Unesco Publishing.

University of Malta. (n.d.). *Research Code of Practice*. L-Università Ta' Malta. Retrieved 29 January 2023, from https://www.um.edu.mt/media/um/docs/research/urec/ResearchCodeofPractice.pdf

Valenzuela, J. P., Bellei, C., & Allende, C. (2016). Measuring systematic long-term trajectories of school effectiveness improvement. *School Effectiveness and School Improvement*, *27*(4), 473–491. https://doi.org/10.1080/09243453.2016.1150861

Veas, A., Benítez, I., Navas, L., & Gilar-Corbí, R. (2020). *A comparative analysis of university entrance examinations using the construct comparability approach*.

Veas, A., Gilar, R., Miñano, P., & Castejón, J. L. (2017). Comparative analysis of academic grades in compulsory secondary education in Spain using statistical techniques. *Educational Studies*, *43*(5), 533–548.

Viennet, R., & Pont, B. (2017). *Education policy implementation: A literature review and proposed framework*.

Vignoles, A., & Dex, S. (2007). Making use of existing data. In *Research Methods in Educational Leadership and Management* (pp. 257–269). SAGE.

Wagner, A. K., Soumerai, S. B., Zhang, F., & Ross-Degnan, D. (2002). Segmented regression analysis of interrupted time series studies in medication use research. *Journal of Clinical Pharmacy and Therapeutics*, *27*(4), 299–309.

Watson, S. L., Watson, W. R., & Reigeluth, C. M. (2008). Systems design for change in education and training. *Handbook of Research on Educational Communications and Technology*, 691–701.

Wei, W. W. (2006). Time series analysis. In *The Oxford Handbook of Quantitative Methods in Psychology: Vol. 2*.

Weimer, D. L., & Vining, A. R. (2017). *Policy analysis: Concepts and practice*. Taylor & Francis.

White, H. (2010). A Contribution to Current Debates in Impact Evaluation. *Evaluation*, *16*(2), 153–164. https://doi.org/10.1177/1356389010361562

White, H. (2020). THE GLOBAL EVIDENCE ARCHITECTURE IN HEALTH AND EDUCATION. *Getting Evidence into Education: Evaluating the Routes to Policy and Practice*.

Whitehurst, G. J. R. (2004). *IPR distinguished public policy lecture series 2003-04—Making education evidence-based: Premises, principles, pragmatics, and politics*. Chicago: Institute for Policy Research, Northwestern University.

Whitney, L., & McIntosh, R. (2001). *An OECD perspective on connecting Educational Research with Policy and Practice in Aotearoa-New Zealand*. Strategic Information Group, Ministry of Education.

World bank. (n.d.). *Education—Learning Outcomes*. Retrieved 15 July 2018, from http://web.worldbank.org/WBSITE/EXTERNAL/TOPICS/EXTEDUCATION/0,,contentMDK:21911176~menuPK:5495844~pagePK:148956~piPK:216618~theSitePK:282386,00.html

Wu, W., Selig, J. P., & Little, T. D. (2013). Longitudinal data analysis. *Oxford Handbook of Quantitative Methods*, *2*, 387–410.

Yin, R. K. (2006). Mixed methods research: Are the methods genuinely integrated or merely parallel. *Research in the Schools*, *13*(1), 41–47.

Zhou, S., Jeong, H., & Green, P. A. (2017). How consistent are the best-known readability equations in estimating the readability of design standards? *IEEE Transactions on Professional Communication*, *60*(1), 97–111.

Appendix A: Average facility and discrimination indices (1999 – 2010)

Subject		1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Social Studies													
	# of Items	74	73	86	85	68	70	64	77	81	79	74	N/A
	F_{avg}	62%	53%	51%	56%	56%	71%	64%	67%	72%	77%	81%	N/A
	D_{avg}	0.42	0.40	0.67	0.80	0.75	0.53	0.51	0.46	0.48	0.44	0.39	N/A
Maltese													
	# of Items	58	65	69	71	66	66	75	75	65	55	61	61
	F_{avg}	72%	50%	45%	54%	53%	66%	64%	77%	62%	67%	76%	80%
	D_{avg}	0.41	0.37	0.65	0.75	0.71	0.49	0.41	0.44	0.44	0.48	0.43	0.31
English													
	# of Items	54	61	79	62	59	45	68	67	68	66	66	61
	F_{avg}	47%	53%	46%	51%	53%	49%	62%	57%	57%	52%	65%	64%
	D_{avg}	0.52	0.35	0.70	0.64	0.67	0.47	0.45	0.49	0.46	0.52	0.50	0.39
Mathematics													
	# of Items	55	57	58	54	59	70	68	62	64	63	60	57
	F_{avg}	50%	53%	47%	55%	55%	58%	61%	66%	62%	64%	62%	66%
	D_{avg}	0.51	0.40	0.62	0.81	0.74	0.54	0.51	0.54	0.54	0.53	0.50	0.50
Religion													
	# of Items	74	73	86	85	84	73	63	60	67	73	73	89
	F_{avg}	62%	53%	51%	56%	57%	79%	61%	78%	70%	78%	76%	77%
	D_{avg}	0.42	0.40	0.67	0.80	0.66	0.54	0.50	0.43	0.36	0.47	0.44	0.32

Appendix B: Action verb word list

List of question verbs taken from Armstrong (2016) and Stanny (2016) and used to determine cognitive levels of the different test items.

Knowledge	Comprehension	Application	Analysis	Evaluation	Synthesis
1	2	3	4	5	6
Arrange	Add	Apply	Analyse	Appraise	Assemble
Choose	Approximate	Calculate	Breakdown	Argue	Adapt
Circle	Articulate	Change	Categorise	Assess	Build
Cite	Associate	Classify	Compare	Award	Collate
Complete the sentence	Assume	Collect	Connect	Conclude	Combine
Count	Clarify	Complete	Contrast	Convince	Compile
Define	Convert	Comply	Debate	Criticize	Compose
Draw	Depict	Compute	Detect	Critique	Construct
Duplicate	Describe	Conjugate	Diagram	Decide	Create
Enumerate	Detail	Demonstrate	Discover	Defend	Deduce
Find	Differentiate	Dramatize	Discriminate	Determine	Delete
Identify	Discuss	Employ	Dissect	Disprove	Derive
Index	Distinguish	Experiment	Examine	Evaluate	Design
Indicate	Divide	Fit	Explain	Feedback	Develop
Label	Estimate	Gather	Infer	Give your opinion	Devise
List	Example	Graph	Inspect	Grade	Elaborate
Locate	Express	Group	Interpret	Judge	Expand
Match	Extend	Illustrate	Maximize	Justify	Extrapolate

Knowledge	Comprehension	Application	Analysis	Evaluation	Synthesis
Meet	Generalise	Interpolate	Minimize	Measure	Formulate
Memorise	Gist	Interview	Point out	Moderate	Generalize
Mention	Give	Manipulate	Question	Perceive	Generate
Name	Give examples	Model	Relate	Prioritize	Hypothesize
Omit	Give reasons	Modify	Research	Prove	Imagine
Order	How	Operate	Review	Rank(order)	Improve
Point	Interact	Perform	Separate	Rate	Instruct
Quote	Multiply	Practice	Simplify	Recommend	Integrate
Read	Note	Predict	Subdivide	Score	Invent
Recall	Observe	Prepare	Survey	Support	Makeup
Recite	Outline	Present	Test	Value	Manage
Recognize	Paraphrase	Produce			Organize
Record	Picture graphically	Schedule			Originate
Repeat	Rearrange	Sketch			Plan
Reproduce	Reorganize	Solve			Prescribe
Select	Rephrase	Translate			Propose
Show	Replace	Use			Reconstruct
Spell	Report	Utilize			Revise
State	Restate	What can be done			Rewrite
Study	Subtract	Work out			Setup
Tell	Visualize				Specify
T/F	Why				Substitute
Tick					Summarize

Knowledge	Comprehension	Application	Analysis	Evaluation	Synthesis
Trace					Suppose
Underline					Synthesize
What					Idea
When					Transform
Where					
Which					
Who					
Write					

Appendix C: Ethics approval

Fwd: Ethical Approval: EDU-2023-02-04T05:58:11-ldsd56

DOUBLESIN, GLENN <glenn.doublesin@durham.ac.uk> 18 February 2023 at 06:03
To: "glenndoublesin@gmail.com" <glenndoublesin@gmail.com>

Sent from Mail for Windows

From: Ethics
Sent: 05 February 2023 15:33
To: DOUBLESIN, GLENN
Cc: HIGGINS, STEVEN E.
Subject: Ethical Approval: EDU-2023-02-04T05:58:11-ldsd56

Please do not reply to this email.

Dear Glenn,

Your supervisor has approved your ethical review form for the following project:

Title: MEASURING CHANGE IN MALTESE EDUCATION: AN IMPACT EVALUATION OF LARGE-SCALE POLICY INTRODUCTION ON LEARNING OUTCOMES AND THE QUALITY OF EDUCATION.;
Supervisor: HIGGINS, STEVEN E.;
Expected Start Date: 30 September 2016;
Application Reference: EDU-2023-02-04T05:58:11-ldsd56.

Based on your responses your project has been categorised as (ethically) low risk and no further review is required before you start work.

Please be aware that if you make any significant changes to your project which mean that ethical approval may be required, you should complete and submit a revised ethical review form.

If you have any queries relating to the ethical review process or requirements for review, please contact your supervisor in the first instance. If you have any queries relating to the online system, please contact research.policy@durham.ac.uk.

1 of 1 18/02/2023, 06:06