Durham E-Theses

The development of an assessment to identify deficits in facial expression decoding in young children

BAILEY, KATHARINE,ELIZABETH

The development of an assessment to identify deficits in facial expression decoding in young children

Katharine Bailey

April 2011

Abstract

The development of an assessment to identify deficits in facial expression decoding in young children

Katharine Elizabeth Bailey

Emotional intelligence (EI) has been found to relate to positive outcomes not only in personal and social development but also in academic achievement. Measurement of EI to identify deficits at an early stage presents opportunities for remediation for those at risk of under-achievement. There are several instruments that claim to measure EI and the more convincing of these do so by capturing the proficiency of individuals in specific abilities. One of the abilities often explored is the decoding of emotional facial expressions. Examination of a group of important EI instruments found that, in all cases, they were mediated by language to some degree. The language issue has implications for valid measurement of children with low vocabulary skills, English as an additional language and those with learning difficulties. The instruments reviewed were often complex and time-consuming to deliver and not appropriate for use by non-specialists. Taken together, these factors limited their application to research with little scope for practical use in the classroom.

The thesis describes the development of FACES, a new test of EI that identified deficits in facial expression decoding in young children. In two studies, an instrument was developed using the Rasch model and was found to be valid and as reliable as the most widely used existing measures. Importantly, the scale used an innovative paradigm, which reduced the use of language and involved children in the development of the items to ensure it was accessible and enjoyable. The scale was found to quickly and successfully identify a low achieving at an age that allowed for intervention if appropriate. Overall, the findings suggested that the FACES measure of EI was suitable for use in the classroom as a quick and simple screening instrument that could contribute to an holistic profile of information on pupils, helping teachers identify those at risk.

The development of an assessment to identify deficits in facial expression decoding in young children

Katharine Elizabeth Bailey

Master of Research (Education)

School of Education

Durham University

2011

Table of Contents

List of figures

List of tables

Nomenclature

Acronym	Definition
CADATS	Computer Assisted Design Analysis and Testing System
CEM	Centre for Evaluation & Monitoring
CTT	Classical test theory
DANVA	Diagnostic Test of Nonverbal Accuracy
EI	Emotional intelligence
EKT	Emotional Knowledge Test
EMT	Emotion Matching Task
EQ-i	Emotional Quotient Inventory
EYFS	Early Years Foundation Stage
fMRI	Functional magnetic resonance imaging
IRF	Item response function
IRT	Item response theory
JACBART	Japanese and Caucasian Brief Affect Recognition Test
LEAS	Levels of Emotional Awareness Scale
MEIA	Multidimensional Emotional Intelligence Assessment
MSCEIT	Mayer-Salovey-Caruso Emotional Intelligence Scale
MSCEIT-YV	Mayer-Salovey-Caruso Emotional Intelligence Scale – Youth Version
PCA	Principal components analysis
PIPS	Performance Indicators in Primary Schools
PIPS BLA	Performance Indicators in Primary Schools Baseline Assessment
PONS	Profile of Nonverbal Sensitivity
PSED	Personal, social and emotional development
QCA	Qualifications and Curriculum Authority
SREIT	Self-report Emotional Intelligence Test

x

Acknowledgements

I would like to thank the following;

Dedication

This thesis is dedicated to Grant, for his unfailing love and support, Charlotte and Georgia, for being truly excellent guinea pigs, Mummy, for always being there, Daddy, for giving me good advice and keeping me on track and Sarah for keeping me grounded.

1 Introduction

1.1 Historical perspective

The focus of psychological research has shifted and evolved over the years. In the early twentieth century much work centred on examining overt phenomena which could be observed and recorded. This behaviourist paradigm included the work of many important psychologists including John Watson and Frederick Skinner. They focused on the need for scientific rigour in research and employed methods of objective and verifiable measurement. To Watson, Skinner and many of their contemporaries, emotion was not only considered irrelevant and without function, but indeed, was thought to interfere with their investigations.

In parallel with the work of behaviourists, other early psychologists such as William James began to think in terms of using introspection to understand more about people's lived experiences and examine unconscious phenomena. The revolutionary work of Charles Darwin writing in the nineteenth century had suggested that emotions served a purpose. He highlighted the evolutionary advantage of the physiological changes that resulted from emotions, such as the 'fight and flight' reflex. James himself had a particular interest in the theory of emotion and, like Darwin, suggested that emotion was connected to physiognomy.

It is mainly in these early foundations of introspection that research into emotion and its function is rooted. Despite some interest among early psychologists, research into emotion remained in the background of the discipline until fairly recently. The growing interest can be largely explained by an increasing diversity of methods to study emotion more objectively and scientifically. For example, using fMRI scanning, it is now possible to identify neural correlates of specific emotion-related experiences in individuals. Another contributing factor to the increased interest is the extent to which, as humans, we are interested in emotions, personality and psychology. Since recent influential work was published on emotional intelligence (Salovey and Mayer, 1989), there has been a huge growth in interest from the general public and much of this can be traced back to the publication of Daniel Goleman's popular science book 'Emotional Intelligence' (Goleman, 1995). To some extent, this has diverted attention from some interesting, definable psychological concepts by extending the reach of emotional intelligence to cover a whole range of academic, personal, social and career-based competencies.

Research into emotion has permeated many disciplines within psychology. Evolutionary psychologists look for functional explanations for emotional behaviours; biological psychology investigates the relationship between emotion and physiologies as controlled by the autonomic nervous system for example, sweat reaction when an individual is anxious; personality researchers are interested in the genetic contribution of emotion; social psychology explores how emotional experiences can be influenced by collective ideas

and beliefs and psychoanalysts look to accessing subjective emotional information to facilitate personal change.

While acknowledging the breadth of the topic, this study will be carried out from a cognitive perspective and particularly from within the body of research on emotional intelligence (EI). It is concerned with identifying aspects of behaviour and personality that are related to successful outcomes in important aspects of people's lives. Much interest in this field has centred on the extent to which individual differences in cognitive ability can predict success in areas such as academic achievement, careers, and personal, social and emotional relationships. Studies have found cognitive ability, as measured by tests of mental ability, to account for between 10% and 25% of the variance in performance related outcomes (Cherniss, 2004). This has led researchers to look for other abilities or 'intelligences' that may explain more of that variance and so contribute to improved performance. The study of emotional or social competencies is one of those areas and it is from this approach that this research is situated.

1.2   Theoretical perspective

1.2.1 Emotional intelligence

The exact definition of emotional intelligence is still a matter for debate, it being a relatively new concept. The term is used in many different ways. The simplest and, arguably most consensual definition, is that it refers to the specific ability to manage emotional information (Mayer et al., 2008). Studies differ on what they consider to be 'information'. The research literature discusses a range of emotional information including one's own feelings and those of others, visual and auditory cues, verbal information and physiological responses.

Mayer *et al* provide a more detailed, and now widely accepted definition. Emotional intelligence is the ability to understand and to problem-solve that involves:

- Managing emotional responses
- Understanding emotions and emotional meanings
- Appraising emotions from situations
- Using emotion for reasoning
- Identifying emotion in faces, voices, postures and other content

In essence, the authors claim that emotional intelligence refers to an individual's capacity to process affective information from their surrounding environment and use this information to adapt the way they interact with that environment. Behind this definition is an underlying assumption that individuals will differ in their ability to perceive, understand and utilise that information. A further assumption is that these differences in ability have a substantial contribution to make to individuals' intellectual and social well being. It is these assumptions that have provided much of the impetus into emotional intelligence research.

There is a general consensus on this definition in much emotions research although it is far from being uncontentious. Elements of this definition are appealing in that it is clear how the EI ability can be demonstrated. Individuals' ability to identify emotions is accessible empirically. To some degree this may also apply to the ability to appraise emotions from situations and understand emotional meanings. However, it is unclear how this definition would explain or demonstrate individuals' ability to manage emotions or use emotions for reasoning. There is also the danger of creating a circular argument from the assumption that individual differences in EI contribute to intellectual and social well-being. If research assumes EI will contribute to success, it is likely success will be linked to EI whether or not there is a sound basis for that assumption. Despite these criticisms, Mayer and colleague's EI definition is well established and forms a framework for much research in the area making it appealing as a foundation for further work.

1.2.2 Models of emotional intelligence

Researchers have conceptualised emotional intelligence in different ways. Mayer *et al* (2008) provide a helpful distinction with which to discuss the different models. They divide the approaches according to whether they deal with emotional intelligence as a series of specific abilities or as a more global ability. A third approach deals with emotional intelligence from a mixed-model perspective.

The first group approaches the components of emotional intelligence as a set of fundamental and discrete skills. This is known as the specific-ability approach. Much of focus in this area has been on the accuracy of perception of emotion and is rooted in research into nonverbal perception and communication. Nowicki and Duke's model, for example, studied accuracy in perceiving emotion in adult and child faces and also through voice and posture (Nowicki and Duke, 1994). Other models have looked at the use of emotional information and how it affects thinking. For example, how positive emotions allow people to perform better in the workplace (Boehm and Lyubomirsky, 2008). Other work has included looking at how emotional intelligence impacts on reasoning and people's management of their own emotions.

A second group deals with models which examine the components of emotional intelligence holistically. These are commonly referred to as integrative approaches. Mayer and colleagues have contributed significantly to the field in studying emotional intelligence from this perspective (Mayer et al., 2004). Their model consists of four branches of emotional intelligence; (a) accurately perceiving emotion, (b) using emotions to facilitate thought, (c) understanding emotion and (d) managing emotions. The branches are ordered according to their level of involvement with other major psychological sub-systems such as goals, self knowledge and social awareness. The first branch is very low level and deals mainly with perceptual and expressive processes. The second branch involves incorporating perceptions and expressions to assist thinking. The third reflects the ability to analyse emotions and understand their outcomes. The fourth and final branch deals with

management of emotion which incorporates elements of the wider personality, for example, to reframe how we think of a situation in order to think more positively. Mayer *et al* (2004) present a compelling argument that each of the four branches of their model is seen to develop from early childhood, with each branch preceding the next. Another useful model in this second group is Izard's theory of differential emotions (Izard, 2001). Izard reflects the integrative approach by combining emotion recognition with interpretation of a situation. This model is particularly useful as Izard has developed instruments using age appropriate materials to study how emotional intelligence works in younger children. Interestingly, it is possible to see how specific ability approaches can play a part in the integrative models as both Mayer *et al* and Izard's integrative models include perceptual tasks.

The final group deals with research which addresses emotional intelligence from a mixed-model perspective. These approaches are characterized by their broad definitions which tend to include some of the properties of emotional intelligence as described above but then draw in other elements of social competence and skill. These models include the work of Bar-On who has been influential in the field and will be discussed later.

The integrative approaches are convincing for two reasons. Firstly, Mayer *et al* and Izard make explicit the components of their respective models and clearly describe the interplay between these components. Secondly, each of these components represents an ability in individuals that can be measured. This gives their theoretical models a real-life application. Mixed model approaches have been criticized as lacking definition. Mayer *et al* (2008) suggest that the mixed-model approaches have extended the findings from integrative models and, in the process, lost the primary focus of emotional intelligence. The simplicity of the specific ability approaches is convincing. They address particular aspects of emotional intelligence and relate these to particular outcomes or behaviours which lay firm foundations for identifying those at risk and developing remediation.

Consideration of the different approaches suggests that the specific ability models are the most appealing as a basis for further research: they are easy to operationalise and most open to empirical support. The simplicity of addressing specific aspects of EI is attractive and yet there is scope beyond the specific abilities approach, as the integrative approach also considers discrete skills and could build on any development of discrete models. Having explored the different approaches to EI, this study will now consider ways in which EI can be measured.

Tapping into emotional intelligence in an objective way raises the same issues that come with any study of covert phenomenon. It is problematic to make observations of emotion-related processes and many methods often rely on individuals' reporting of inner experiences which can be confounded by many factors including language, motives and beliefs. Mayer *et al*'s hierarchical approach discussed earlier presents a convincing developmental model which, in the initial stages, addresses simple processes, one of these being the perception of emotion in others. Perhaps it is possible, then, to measure

emotional intelligence using simple perceptual tasks? This would require minimal use of language and would lend itself to simple and appealing formats suitable for use with young children. A considerable body of research has considered facial expression perception specifically.

1.2.3 The role of facial expression decoding in emotions theory

The specific ability and integrative models of emotional intelligence all incorporate the perception of emotion in others. How each model addresses this ability differs. Some models incorporate several modalities for emotion perception including facial expression, posture, voice and gestures. Others focus on single abilities of which facial expression decoding commonly appears in the literature.

Most models of emotion will feature facial expression whether explicitly or implicitly, although causes and function may differ between theorists. Research largely converges on a consensus that facial expressions are produced as a result of our appraisal of an emotional situation. Where theorists differ is in the variety and complexity of the emotions involved.

Research into the function of facial expression decoding is largely rooted in a particular approach to emotion which posits the existence of a set of discrete emotions which underlie all emotional experience. This approach is given the term 'basic emotions'. A useful analogy to describe this approach is provided by Yiend and Mackintosh (2005). They compare the processing of colour by the visual system with the processing of emotional information. The full spectrum of colour is represented in the brain by the stimulation on the retina of just three types of cones. Basic emotions theory suggests that a full range of emotional experience can be explained by a combination of a small set of discrete emotions. For example, research has suggested that the basic emotions of joy and acceptance combine to produce the more complex emotion of friendliness (Plutchik, 1962). The basic emotions tradition encompasses the work of many researchers and, as a result, there is some dispute around the number and type of emotions proposed (Power and Dalgleish, 1997). Arnold's model proposes 11 basic emotions; anger, aversion, courage, dejection, desire, despair, fear, hate, hope, love and sadness. Weiner and Graham's model proposes just two – happiness and sadness. The basic emotions proposed by the key emotion theorists are summarised in Table 1.

Table 1: Basic emotion theorists and the emotions they propose

Emotion theorist	Fundamental emotion
Arnold	Anger, aversion, courage, dejection, desire, despair, fear, hate, hope, love, sadness
Ekman, Friesen, and Ellsworth	Anger, disgust, fear, joy, sadness, surprise
Frijda	Desire, happiness, interest, surprise, wonder, sorrow
Gray	Rage and terror, anxiety, joy
Izard	Anger, contempt, disgust, distress, fear, guilt, interest, joy, shame, surprise
James	Fear, grief, love, rage
McDougall	Anger, disgust, elation, fear, subjection, tender-emotion, wonder
Mowrer	Pain, pleasure
Oatley and Johnson-Laird	Anger, disgusts, anxiety, happiness, sadness
Panksepp	Expectancy, fear, rage, panic
Plutchik	Acceptance, anger, anticipation, disgust, joy, fear, sadness, surprise
Tomkins	Anger, interest, contempt, disgust, distress, fear, joy, shame, surprise
Watson	Fear, love, rage
Weiner and Graham	Happiness, sadness

Source: Power and Dalgleish, 1997

Despite the affordance given to this body of work, there remain challenges for basic emotions theory. A key problem is that if there are indeed basic emotions, why is there so much disagreement about how many there are and what those emotions are? Theorists suggest that basic emotions can be combined to explain the full range of emotions, but how they explain this interrelatedness and the complexity of the processes involved remains largely unanswered. A further issue is that, of the studies referred to in Table 1, some consist of scientific and empirical work but some of the research is speculative and the extent to which generalisations can be made is limited.

Another influential approach is that of the 'dimensionality' of emotion which goes some way to overcome this difficulty. Where basic emotions theory suggests a set of discrete emotions which, when combined, explain the full range of emotionality, the dimensional approach explains emotions as being locations within a two-dimensional space. The Positive and Negative Affect Schedule (Watson et al., 1988) was developed by studying physiological responses to emotional material. This, and other dimensional models, consist mainly of two dimensions (occasionally three) and can be conceptualized as a grid. In the example in Figure 1 the dimensions are 'arousal' and 'valence'. Arousal indicates the level of response experienced and valence, whether that response is positive or negative. This model goes someway to explaining the wide range and intensity of emotional experience while accommodating the emotion labels which are part of everyday life. For example, sadness might be located in the bottom left hand box in Figure 1 – a combination of low arousal and negative affect. However, there remain the problems of identifying how many dimensions and what those dimensions might be.

6
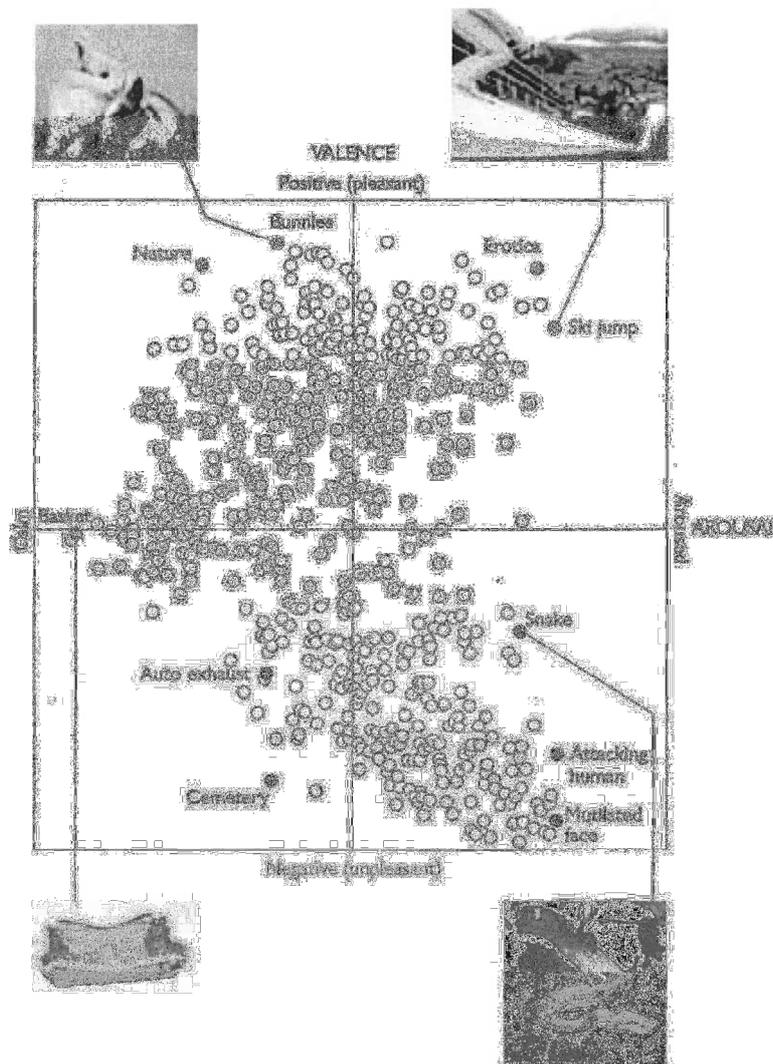
Figure 1: The affect grid

Source: (Dawson et al., 1999)

1.2.4 The importance of facial expression decoding

The perception and decoding of facial expressions is considered to be an essential skill to enable humans to function effectively in a highly complex social environment (Montagne et al., 2005). Being able to quickly and accurately determine the emotions of others through facial cues allows individuals to react appropriately and to modify their own behaviour to effect the outcomes of social interactions. Heise (1985)suggested that, although language is crucial to social functioning, only facial displays are able to fully and accurately communicate the range of human emotional information. This is particularly pertinent when spoken language cannot be relied upon. He illustrated this with a useful example. He posited that, on visiting a foreign culture, you might demonstrate a particular behaviour which results in a native of that country emitting an expression of horror. The ability to

accurately interpret this facial expression would lead you to modify this behaviour and generalise that that particular behaviour was not acceptable in that country. In this way, humans are able to be adaptive in social situations. Kemper (1981) expanded on the importance of emotion decoding. He argued that not only are our faces our primary channel of communicating emotional information but that facial emotional displays provide important evidence to us of others' perception of their own power and status and how that changes during the social intercourse.

From the moment babies are born, they begin to develop their skill in attending to, and decoding, human facial expressions. Bowlby's seminal work on attachment theory put much weight on the importance of parent-infant facial communication as a precursor to appropriate social functioning (Bowlby, 1999). From around the age of 12 months, it is believed that infants begin to use information from the decoding of facial expressions to moderate their own behaviour. This process, known as social referencing, relates to how infants take their cues for appropriate behaviours and emotions from their interactions with, and observations of others around them (Hertenstein and Campos, 2004). As the skill develops children will continue to use the information to modify behaviours and increasingly to facilitate their interactions with others. Research has suggested that maturation of the facial decoding skill is complete by the age of around 8 (Baron-Cohen et al., 1996). Herba and Phillips (2004)suggest that mastery is not achieved until adolescence or early adulthood, particularly in regard to speed of processing. They suggest that the skill is mediated by other factors such as sex, socio-economic status and verbal ability. The effect of mediating factors and the suggested period of development are interesting when considering the importance of deficits and the potential for, and impact of interventions.

1.2.5 Universality and facial expression decoding

Facial expressions allow people to perceive emotions in others and convey their own feelings in return. Charles Darwin studied facial expressions from an evolutionary perspective by comparing the expressions of humans and animals (Darwin, 1998). His suggestion was that facial expressions served to convey emotional status to others. A well cited example he used was that of a wolf snarling. This expression would be emitted to ward off potential adversaries. The basic components of a snarl (exposed teeth and drawn up lip) can still be observed in contemporary human facial activity. The key criticism of Darwin's work in this area is that he suggested that the expressions of humans no longer served a contemporary function and were merely an artefact of our distant ancestors. Yiend and Mackintosh (2005) suggest that to some extent this may be correct but that humans have evolved to control and manipulate the production and recognition of facial expressions to allow them to function in a highly complex social environment. For example, individuals portray expressions that indicate emotional states they do not feel in order to deceive and manipulate. There is also a considerable body of comparative research which has explored facial behaviour in primates and how that relates to human behaviour. A

recent study (Parr et al., 2008) found that, in decoding, chimpanzees and humans attend to the same configurations and movements of facial expression (for example, degree of mouth closure) which adds weight to Darwin's argument for an evolutionary function.

A vast body of work exists on the evidence for universality in emotional facial expressions. That is, that the mechanisms for facial expression decoding, and the expressions themselves, are the same in all humans (and, to a certain extent, in non-human primates) with the same genetic basis. Paul Ekman is responsible for a great deal of this research and, indeed, it is the findings of his seminal study that have formed the basis for much work since (Ekman, 1973). In his research, American college students were asked to produce facial expressions appropriate for a number of emotions. These were photographed and taken to a remote area of Papua New Guinea where inhabitants had little or no experience of the wider world. Indigenous individuals were asked to identify the emotions that they saw portrayed by the students in the photographs. Ekman also performed the experiment in reverse, recording the New Guineans' facial expressions for the students back in the USA to decode. There was a high level of agreement for happy, sad, anger, fear, surprise and disgust expressions[1]. Evidence from this, and similar studies, has been used to propose the genetic basis for the production and recognition of facial expressions, termed the 'universality' position. This assumes that, because the New Guineans were largely isolated from wider societies, yet produced and recognised the same expressions, there was an innate basis for the ability. A more recent study used photographs of the emotional displays of athletes from 35 different countries. The authors found a high level of consistency in the production of initial emotional expressions providing further support for the universality approach. Interestingly, they also found evidence of the cultural effect on emotional expression production as, after initial emotional expression display, subsequent display was moderated by the culture of the individual (Matsumoto et al., 2009).

Psychologists often work with newborns and very young infants to identify phenomenon with a genetic basis because they have had very little opportunity to learn from their environment. Research has shown some consistency in the production and recognition of faces by infants (Ekman and Oster, 1979, Nelson, 1987, Farroni, 2007). Medicus *et al* found that children blind from birth exhibited facial expressions which could be identified as belonging to basic emotion categories, adding further weight to the argument for universality (Medicus, 1994). One compelling study compared a cross-cultural sample of congenitally and non-congenitally blind athletes and found that production of spontaneous facial expressions of emotion were consistent across sightedness and cultures (Matsumoto and Willingham, 2009).

Despite the universality theory being the dominant perspective in the field, Russell (1995) summarised key challenges. Central to the theory of Ekman and others is that facial

---

[1] The emotion terms used here, and throughout this thesis, are shown as referred to in the original studies, while acknowledging that the terms are not necessarily grammatically consistent.

expressions are linked to underlying emotions. This, as Russell points out, "*presupposes that emotions are in the face to be recognised*". It has not been proven that, across the world, happy people smile and angry people frown. Measuring underlying subjective experiences is highly problematic, particularly when the aim is to make generalisations. Russell's argument, then, is that universality addresses merely a series of facial movements that are similar but that does not necessarily suggest any universality in the emotions. So strong is the evidence for the 'universality' approach, however, that Russell presented no opposing theory. Merely, he suggests that further research goes back to the fundamentals of facial expressions to generate new theory. If we make an assumption that facial expressions map onto underlying emotions, there still remain challenges to Ekman's work and the extent to which it can be used in contemporary research.

The universality theory relies heavily on the use of language to gather evidence. For example, in Ekman's study, people were asked to produce an expression appropriate to a spoken word or phrase. This raises methodological issues. How were the words and phrases chosen? And how can we be sure that responses are not affected by the understanding of language? It is likely that there were cultural differences in the interpretation of verbal cues. When participants were asked to identify expressions, the stimuli were photographs of heads. These disembodied faces are but one piece of information that people may use to identify emotions. They may also use head position, hand and body gestures and tone of voice. Finally, Ekman's comparison of cultures was based on a large sample of western literate people and much smaller samples of isolated, illiterate communities.

Although Ekman's work represents a substantial proportion of work in the area, others have replicated, and extended his findings. Baron-Cohen *et al* (1996) conducted a study on facial expressions that extrapolated Ekman's research beyond the six basic emotions. Baron-Cohen supported the existence of a small set of emotion expressions but suggested the existence of a further set of universally recognisable expressions of cognitive states such as distrust and recognition. His study involved adult participants from Britain, Japan and Spain. Baron-Cohen found strong evidence for universality in these cognitive expressions: recognize, threaten, regret, astonished, worried, distrust, contempt and revenge. He repeated the experiment with 80 British school children aged between 8 and 11 years with similar results. His hypothesis expected there to be some development in ability between the younger age groups but this was not found and he suggested that facial expression recognition was stable by the age of 8. Baron-Cohen also suggests there is a differential in the rate that skill is acquired for decoding the different expressions but by the age of 5 children can usually decode the basic expressions.

Other research would appear to concur with the idea of maturation of the skill at around this age. Nowicki suggests that four basic emotion expressions (happy, sad, anger, fear) are learned by the age of 10 (Nowicki and Duke, 1994). If these findings are true, it has implications for the development of measurement instruments. Perhaps early identification can pave the way for interventions to improve facial expression decoding skills before they

stabilize? Of course, although identification may be stable by the age of 8, appraisal and manipulation of that information will be developed over a far longer period as individuals become more experienced.

Real life interpretation of facial expressions is complicated by varying degrees of intensity. Someone may be happy or very happy. Typically, humans appear to be able to discriminate between these differing levels of expression. Kuchuk and colleagues (1986) found that three month old babies were able to discriminate between happy faces of differing intensity and consistently attended longer to each level of happy face when presented with a neutral control. A further study found that seven month old infants could identify varying levels of intensity in happy, surprise and fear expressions (Ludemann and Nelson, 1988). This evidence is interesting as it has implications for the development of stimulus material for testing the facial expression decoding skill.

1.3   Application for measures of emotional intelligence

The body of research to this point has sought to explore facial expressions for their own sake. Other research has looked at how individual differences in ability to decode emotion expressions have been used to predict wider behaviours. This suggests such an approach could be used to differentiate ability. Marsh *et al* (2007) conducted an experiment that linked accurate identification of the fear expression with pro-social behaviour (behaviours that improve outcomes for others). Marsh *et al* highlighted previous work which suggested that seeing a fear face in another generated empathy which was necessary for moral socialization. Using this as a basis, they hypothesised that ability to decode the fear expression should predict pro-social behaviour. They used a standard test of expression recognition to measure for individual differences in ability and also presented each participant with a task designed to elicit a degree of pro-social behaviour, such as donating money to someone in need. They found that there was a tendency to pro-social behaviour associated with the ability to interpret the fear expression.

Of course, caution should be used in the interpretation of correlational studies. The fact that there is a relationship does not suggest that it is necessarily causal. This is particularly important where children are concerned, as results may be used to inappropriately 'label' a child. It should also be considered that, once again, the study relied on language, and although the authors describe measures to limit the effects of individual differences in language, they cannot be ignored.

The literature shows that deficits in emotional understanding have been found in people with autistic spectrum disorders. Children with autism have been shown to have difficulties in matching modes of emotional expression, including facial expressions (Hobson, 1986) and people with Asperger syndrome have been found to exert more intellectual effort to process facial expressions (Attwood, 2000). Using a new test, Golan *et al* (2006) compared the ability to discriminate emotions between individuals with Asperger syndrome and those in a control group. Golan points out that, unlike those with autism, individuals with

Asperger syndrome can recognise the six basic emotions but have difficulty in recognising more complex emotions. As such, standard tests of facial expression recognition fail to identify Asperger syndrome in many individuals. Golan added new dimensions to a standard test, including voice and moving images of faces. His study found that participants with Asperger syndrome were differentiated in the new test.

At this point, it is important to justify and qualify the use of the term 'deficit' within this thesis. It is not always helpful, and may be potentially damaging, to refer to deficits within this context. A child with a low score in EI may certainly have concomitant difficulties with socialising although the outcomes for children may be vastly different. A child with low EI score but who is otherwise high ability and from a good home background may not be a cause for concern. On the other hand, a child with a low EI score who is generally low ability and who has a deprived home background is almost certainly more likely to benefit from some degree of intervention. It would be more helpful, then, to think of an EI measure as adding to a profile of information on a pupil which might include cognitive ability, academic achievement, personal, social and emotional development, social and cultural factors and home background. While acknowledging this complexity and reinforcing the need for an holistic approach to identifying those in need of help, the use of the term 'deficit' will be maintained as a helpful term for the purposes of this research.

If ability to decode facial expressions is to be used to identify deficits then it is important that the test discriminates accurately. Golan's test goes some way to providing a more ecologically valid assessment than those that rely solely on 2-dimensional images. However, it is not at all clear that an assessment such as this would be suitable for children: the test required 45 minutes of consistent concentration. A common weakness in psychological research is to employ tools and instruments with adults in mind and then adapt them for use with children. A tool that can be used with young children should consider the children's strengths and weaknesses from the early stages of their development.

1.4   Educational relevance

Were a test to be developed that could accurately and objectively measure children's emotional intelligence, what bearing would this have in an educational context? In recent years there has been an increased interest in the personal, social and emotional development of children within education. If facial expression decoding is found to correlate with emotional intelligence more generally, it could be used as a basis for an objective test to identify those with deficits. Were those deficits to be remediated, it may be possible to improve outcomes for individuals.

In England, social and emotional factors have been introduced into the curriculum over a 20 year period as the emphasis on early childhood education has shifted (Kwon, 2002). The Education Reform Act 1988 introduced the National Curriculum which clearly set out areas of learning for children in compulsory schooling. Prior to that, there was little intervention from central government and no formal regulation of curriculum delivery. The National

Curriculum focused on raising standards across the country by providing a structured curriculum in academic subjects. No reference was made to personal or social skills or outcomes. In the 1990s, early years' education was specifically addressed by the introduction of Desirable Outcomes for Children's Learning (1996). This set out skills, knowledge and ability considered to be important precursors to compulsory education. The Desirable Outcomes included physical development, knowledge and understanding of the world, and personal and social development alongside language, literacy and mathematics. The Qualifications and Curriculum Authority replaced the Desirable Outcomes with the Early Learning Goals (1999) extending the scope to include the personal, social and emotional development of children in the first year of compulsory education. This was formalised by the Early Years Foundation Stage (EYFS) guidance (DCSF, 2008) for use in England from September 2008 onwards which further emphasised the importance of personal, social and emotional development (PSED).

The statutory assessment of the Early Learning Goals, that form the backbone of the current framework, aims to measure progress in areas including disposition and attitude, social development and emotional development. The rationale behind this is to promote a positive sense of self, a positive disposition to learn and emotional well-being.

It is clear that the development of emotional skills will contribute to personal and social well-being. Is there any evidence to support a link between emotional intelligence and academic achievement?

Studies have found links between emotional intelligence and school performance. Trentacosta and Izard (2007) found that teachers may share closer relationships with students who have a wide range of emotion competence. They also suggested that emotion regulation predicted academic competence in young children, although teacher rating contributed to half the academic measure. Parker *et al* (2004) compared the grade point average of high, middle and low achieving high school students and found academic success to be significantly related to a measure of EI. Similarly, Schutte *et al* (1998) found measures of incoming college students' EI to predict their end of year grade point average. Of course, there are issues with direction of causation here: those students doing well academically may be more likely to report happiness. Also, both studies utilised self-report which is often mediated by other factors such as personality and vocabulary. Indeed, EI as measured by self-report has been found to correlate weakly with performance measures (Brackett et al., 2006) and self-report would not be an appropriate measure for young children. Identifying deficits at an early stage would be important if remediation were to be appropriate.

Further studies have suggested some links between socio-emotional skill and cognitive function. Kohn & Rosman (1973) used instruments of social-emotional function to predict cognitive functioning and found an association between measures of apathy and withdrawal and poor cognitive functioning in pre-school children. Miles and Stipek (2006) found significant associations between social skills (aggression and pro-social behaviour)

and literacy in children from low-income backgrounds at particular risk of negative outcomes. Marsh *et al* (2007) found that individuals who were able to recognise fear expressions behaved more pro-socially than those who were not. In a comprehensive review, Blair (2002) addressed the functional role of social and emotional skills in cognition from a neurobiological perspective. His work converged on there being a significant contribution from emotion in organising and directing cognition. For example, deficits in strategic thinking have been associated with poor attributions of the self as a learner.

The studies examined here are helpful in terms of making general links between EI and other outcomes with the assumption being that good levels of EI are going to benefit everyone. While this might be appropriate for the majority of children, caution should be used. Arguably, not everyone may benefit from the identification, and possible remediation, that might result from this assumption. An 'emotionally unintelligent' but high achieving child may be turned from an original thinker into an average pupil. A child with low EI and from a deprived background may be protected by being emotionally 'switched off' and remediation may be more damaging in the long term.

Bearing in mind the sensitivities around making inappropriate generalisations, the body of research suggests that emotional intelligence may predict cognitive and academic as well as social outcomes. This is appealing for the development of a test to identify deficits in EI that would be used as part of a holistic approach to identifying pupils at risk.

1.5 Measuring emotional intelligence

There are a number of tools that claim to measure emotional intelligence. In their useful review, Mayer and colleagues outlined the measures of EI that related to their categorisation of EI models. They contrasted these key instruments against a number of criteria adapted from Standards for Educational and Psychological Testing (Mayer et al., 2008). The measures were grouped into specific ability, integrative and mixed models as described earlier and are shown in the Table 2.

Table 2: Emotional intelligence measures

Test	Description
Specific ability measures	
DANVA **Diagnostic Test of Nonverbal Accuracy**	Multiple choice responses to indicate which of four emotions (happy, sad, angry, fearful) is present in three types of stimuli; facial expressions, voice and posture
JACBART **Japanese and Caucasian Brief Affect Recognition Test**	Seven emotions (happiness, contempt, disgust, sadness, anger, surprise and fear) in Japanese and Caucasian faces are portrayed in video format with test-taker asked to correctly identify the emotion present.
LEAS **Levels of Emotional Awareness Scale**	Test-taker is presented with twenty vignettes involving 'you' and one other individual. After reading the vignette, he or she is asked how they would feel and how the other person would feel. Scoring is on a continuum of low to high emotional awareness.
Integrative model measures	
EKT **Emotional knowledge test**	Series of evolving tests including EMT (Emotion Matching Task) consisting of four parts which measure receptive and expressive emotional knowledge, emotion situation knowledge and emotion expression matching. The matching task involves children matching facial expressions with situations or causes.
MSCEIT **Mayer-Salovey Caruso Emotional Intelligence Scale**	Eight tasks are presented utilising different item types and response scales. The areas measured are perception in faces and landscapes, using emotions in synaesthesia and facilitating thought, understand emotional changes over time and blends and managing emotions in oneself and relationships.
Mixed model measures	
EQ-i **Emotional Quotient Inventory**	Test takers are presented with 133 items that require self-judgment responses that cover five factors; intrapersonal, interpersonal, adaptation, stress management and general mood.
SREIT **Self-report Emotional Intelligence Test**	A 33 item self-report inventory giving an overall EI measure.
MEIA **Multidimensional Emotional Intelligence Assessment**	A self-report inventory with 118 items over 10 scales.

Source: (Mayer et al., 2008)

A review of the literature revealed that the instruments Mayer and colleagues investigated did not constitute an exhaustive list. In fact, several other tests are referred to in the emotion literature including the PONS (Profile of Nonverbal Sensitivity) test, (Rosenthal et al., 1979) and the Pictures of Facial Affect (Ekman and Friesen, 1976). Many others were less well reported in the literature and, importantly, were developed from within a different perspective and conceptual framework making it difficult to evaluate them against those purporting to measure EI as described thus far. One notable exclusion from Mayer and colleagues' evaluation was the MSCEIT-YV, a youth version of the MSCEIT described in Table 2 above. This assessment had considerable appeal in that it extended one of the most widely used instruments to a child and adolescent population. However, at the time of

writing, its test properties had not yet been published which meant it could not be evaluated in the same detail as the other measures.

To discuss the instruments, the authors evaluated them against a set of broad criteria which were grouped into three categories:

- Sound test design as specified by validity and reliability
- Structure of EI measurement
- Convergent validity

The first category considered whether the instrument had a sound test design. This involved examining several criteria which are shown in the Table 3:

Table 3: Criteria for judging validity and reliability of an instrument

Criteria	Description
content validity	the extent to which emotional intelligence as a concept is represented by the measure i.e. whether the instrument is measuring the construct as defined by the EI conceptual model
response-process evidence of validity	measures of ability require an assurance that the test-taker is presented with a question that allows them to demonstrate their ability in the construct being measured and elicits a response that can be judged for correctness
reliability	the consistency with which the test measures different people and measures individuals over time

The second category addressed the structure of the EI measurement and whether this was measuring one EI ability or a range of abilities. If EI is to be considered as the unified intelligence its proponents suggest, then findings should identify a hierarchy of factors and subscales.

The final category was intended to examine convergent validity or the extent to which the measures correlate with other measures of EI.

In their findings, the authors suggested that, in general, the specific ability and integrative models provided reasonable evidence of adequate test design with good content validity, appropriate scoring methods and reliabilities ranging from $r = 0.80$ to $r = 0.92$. Evidence of a general EI measure with specific factors was found, particularly in the integrative models. Convergent validity of $r = 0.80$ was found between two of the measures but, overall, low correlations were found and the authors admitted that this was troubling. The mixed models raised issues about the extent to which they were measuring EI. It was argued that these models incorporated other attributes such as assertiveness and flexibility which meant the instruments lacked content evidence for their validity to assess EI. It was also noted that the mixed models relied on self-report which gives self-estimated ability measure, not an actual ability measure. It was posited that these measures were prone to more positive self-judgments, giving scores that did not correlate well with other instruments. Overall, the authors found the specific ability and integrative models more compelling in their claims to measure emotional intelligence than were the mixed models.

Indeed, the authors question whether the mixed models are measuring emotional intelligence at all.

Assuming the remaining specific ability and integrative models are good measures of emotional intelligence, it was necessary to examine which of them could be used with young children. If, as the Baron-Cohen study suggested, the ability to decode expressions is stable by the age of 8, identifying deficits earlier in life would be important if interventions were to be put in place to improve outcomes. It follows that a measure of EI would be useful, if appropriate for children in the first years of full time education, aged 4 to 6.

Five of the models were designed for adults and, although they had been adapted for use with adolescents, the format and content were not intended for the younger age group. However, three of the measures (LEAS-C, DANVA and the Emotion Matching Task (EMT) scale of the EKT) were either suitable for use with children or had separate versions which had been adapted for them (Morgan et al., 2010, Nowicki and Duke, 1994, Bajgar et al., 2005, respectively). These instruments were examined in more detail.

The measures were compared for suitability with younger children using four criteria.

- Reliability and validity
- Scoring method
- Language content
- Stimuli

Firstly, they were contrasted against key criteria laid out by Mayer and colleagues and described earlier. Reliability and validity issues are central concerns in selecting an appropriate instrument and for this reason, content, response-process and convergent validity plus reliability measures were examined. Mayer and colleagues argued that factor structure was also important. Although this is central in arguing for a place for EI as a unified intelligence, it is of less value for evaluating instruments.

Although useful, Mayer and colleagues' criteria were not considered to be thorough enough for a meaningful examination of the relevant issues. In addition, Mayer is the author of one of the instruments evaluated which is likely to colour their interpretation. For these reasons, further criteria were added that were specific to this research.

An often overlooked factor in operationalising a model of emotional intelligence in an instrument is how the correct items are specified. One way is to use expert opinion. In this case, the test designer would pre-specify the correct answer. This assumes that there *is* a correct answer and also that the test designer is correct. A common method employed here is that of judging the correct item on the consensus of a panel of experts. Another method commonly employed with multiple choice items is to allocate the right answer to the choice with the highest percentage of responses. Although an ecologically valid method, this has the disadvantage that the correct answer may change when used with different populations. This can be overcome by using consensus marking over a diverse sample and then setting the correct answers for future use. How correctness is scored will be important when interpreting results and so scoring method was examined. As levels of language

competency would be expected to vary between individuals, it is likely that this would act as a confounding variable in the measurement of their emotional intelligence i.e. children who can better understand the instructions or content and verbalise their response would score higher than those with similar EI but weaker language competency. For this reason, the instruments were also compared for language load. The instruments differed in the type of stimuli used. It is possible that the stimuli used with adults may not yield similar patterns of results and may reflect exposure to different media, for example, preferences for cartoons over photographs. The types of stimuli employed were examined as the final point of comparison.

Evaluation of content validity found that the LEAS-C and EMT measures were well grounded in appropriate theory and the constructs measured reflected that theory well. The DANVA specified that rather than based on theory it was developed using criteria that were considered to represent particular nonverbal behaviours by a norm group. This is good in terms of ecological validity, but causes a problem if the instrument is to be measured against others with good content validity.

All three measures had good evidence of response-process validity with questions reflecting the models they propose. However, the LEAS-C was worryingly dependent on vocabulary which raised the question about whether it was actually measuring EI or vocabulary acquisition.

Convergent validity was low with all measures failing to demonstrate more than weak correlations with other EI measures. The only moderate correlations were found between sub-scales of the same test or with teacher rating scales. Test reliability was good across all the measures.

The DANVA measure was explicit about how correctness was judged using a previously obtained consensus scoring method. Both the LEAS-C and EMT failed to specify how correctness was judged which has implications for interpretation of their findings.

All three measures needed a minimum of verbal ability to ensure access to the test. The DANVA only required the test-taker to be able to identify between emotion labels (for example, happy, sad, etc). The EMT and LEAS-C had a relatively high language load both in presentation of stimuli and, with the LEAS-C, need for use of emotional language in open-ended responses. As far as stimuli were concerned, the tests employed photographs of adults and children.

The DANVA scale would appear to be the most useful instrument for measuring emotional intelligence in young children. The EMT, although meeting most of the criteria for good assessment, was considered to have too high a language load to be able to provide a good measure for young children with lower verbal acquisition and for those with English as a second language. The LEAS-C fell short in a number of areas and, interestingly, is the only one of the instruments to have been adapted from an adult version, rather than developed with children in mind.

Table 4: Emotional intelligence measures appropriate for use with children

Scale	Age range	Content validity	Response-process validity	Convergent validity	Reliability	Scoring method	Language content	Stimuli
DANVA (Diagnostic Analysis of Nonverbal Accuracy)	6 - 10	Based on empirical-normative approach. Denies links with EI theory but based on perceptual accuracy of skills considered important to EI.	Asks test-takers to correctly label photographs with one of four emotions.	Significantly but weakly correlated with other personal and social measurement scores but not compared against other EI scales.	0.88 (coefficient alpha) 0.84 (test-retest)	Correctness judged against previously obtained consensus scoring.	Response judged on correct labelling of emotion perceived.	Photographs of adult and child faces, postures and gestures shown for one second.
LEAS-C (Levels of Emotion Awareness Scale for Children)	9 – 12	Construct reflects theory from developmental levels of emotional awareness (LEA) model – element of EI representing ability to identify and describe emotions of self and others.	Test takers required to describe how target was feeling in sequence of faces and scenes.	Correlated weakly with parental descriptions scale (.01 - .18). Small to moderate correlations (.03 to .3) between sub-scales.	.93, .86 and .89 (inter-rater reliability) 0.71, 0.64 and 0.66 (coefficient alpha)	Criteria for identifying correctness not specified.	High level of language use needed for open ended questioning and for comprehending verbally presented vignettes.	Verbally presented vignettes of between two and four sentences.
EMT (Emotion Matching Task)	3 - 6	Grounded in theory of emotional knowledge which includes ability to recognize emotional expressions, label expressions and understand causes and consequences of emotions.	Matching expressions, labels, situations and causes in series of sub-tests.	Moderate to high correlations between subtests (0.32 – 0.76). Moderate correlations with teacher rating (0.31 - 0.45). No significant correlations found with other measures.	0.88 (coefficient alpha) 0.87 (split half internal)	Method of establishing correctness not made explicit.	Some sub-scales relied on verbal presentation of stimuli (for example, show me the one who got a pretty puppy for a birthday present).	Ethnically diverse photographs of young children's emotional expressions.

1.6   Rasch measurement

When tests are constructed, questions need to be included to discriminate between different levels of ability. This will ensure that inferences about the ability level of all individuals can be made.

Initially the questions will often be judged for level of difficulty by the person constructing the test. Once data has been gathered from test-takers, a facility value (or measure of difficulty) will be assigned to each question. This may take the form of, for example, the percentage of test-takers who answered the question correctly. If 85% of test-takers answered the question correctly, the question would be considered fairly easy and have a high facility value. If only 15% of test-takers answered correctly, the question would be considered very difficult and would have a low facility value. The items taken together would form a test with a score representative of an individual's ability.

1.7   Theories of test development

At this point, there are different approaches to test development. Classical test theory (CTT) is one such approach and is concerned with establishing a true score for an individual through improving the reliability of the test. The assumption of a CTT approach is that the score produced by the test is actually the product of a relationship between the true score (the score obtained if it was a true and consistent reflection of ability) and an element of measurement error. Improving the reliability of the test reduces the error and hence improves its quality as a measurement instrument. This approach has its limitations. The score for an individual is specific to the test from which it was generated which means that comparing different forms of the test becomes problematic. The limited item level information generated does not allow for much improvement to be made to the test. Item Response Theory (IRT) is a more sophisticated approach to psychometric measurement which uses mathematical models not only to create tests and derive scores but also to give sophisticated information at item level to improve the performance of the tests. Fundamental to the IRT approach is that the probability of an individual's response to an item is a function of the ability of the person and the difficulty of the item (IRF – item response function).

In the test design, the items will often be presented sequentially as shown in Figure 2

Q1   Q2   Q3   Q4   Q5   Q6   Q7   Q8   Q9   Q10   Q11

← ———————————————————————————————————— →
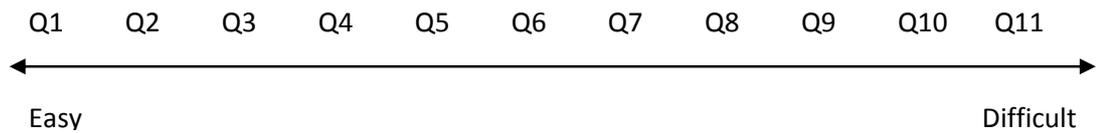
Easy   Difficult

Figure 2: Representation of CTT range of item difficulty

The IRT model would assume that a test-taker who answers questions one to nine correctly is more able than a test-taker who answers questions one to six correctly. This also assumes that a test-taker answering question ten correctly is also able to answer questions one to nine. And that a test-taker answering question eight incorrectly, will answer questions nine to eleven incorrectly.

Measurement systems are widely used and understood as a way of sharing an understanding of, for example, length, volume or capacity. A ruler measures objects in centimetres. A centimetre ruler measures centimetres the same way in every household and classroom across the country. The diameter of a plate measures the same when measured by several different centimetre rulers. Additionally, and crucially, the distance between two centimetres and three centimetres is the same as the distance between seven centimetres and eight centimetres. That is, the centimetre ruler is an equal interval scale. And the same applies for instruments used to measure weight, volume etc. However, when measuring complex human constructs, this concept becomes problematic. Social sciences commonly work with latent traits and self-report that are riddled with extraneous variables that make clear and consistent measurement impossible. Figure 2, above, suggests that the hypothetical test is an equal interval scale but, in reality, the scale probably looks something that that shown in Figure 3. The letters A, B and C, represent three individuals and are placed on the scale at the point where each individual was unable to answer any more questions correctly.

Q1 Q2 Q3   Q4 Q5   Q6   Q7   Q8 Q9   Q10   Q11

← ———————————————————————————————————— →

                                    A   B   C
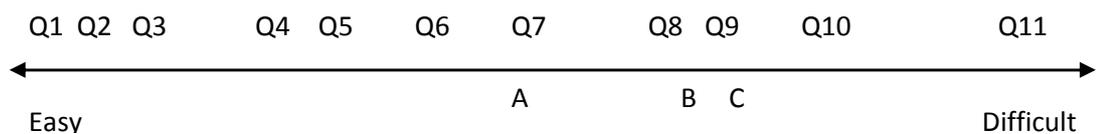
Easy   Difficult

Figure 3: Realistic representation of CTT range of item difficulty

This, more realistic model, shows that although the scores of three test-takers are sequentially ordered, there is a wider ability gap between test-takers A and B than between test-takers B and C. This throws into question the accuracy of the measure and how well it is representing individuals' ability.

Figure 3 highlights the effect that real people and real situations have on the design of a test. IRT deals with this by proposing three different models which are characterised by the number of parameters they use. A three parameter model would account for the probability of an individual getting a correct answer, the discrimination of the item and the effect of an individual guessing. The two parameter model would account for probability of

an individual getting a correct answer and item discrimination. The one parameter model assumes that guessing and item discrimination will not have an effect.

The one parameter IRT model was used as a basis for Danish mathematician and statistician, Georg Rasch, to develop a mathematical model that incorporated the responses of individuals on the test with the facility values of the items (Bond and Fox, 2007). This method forces (statistically) an equal interval scale to provide a more accurate way to measure in the human sciences. As outlined previously the one parameter IRT model has challenges based on the assumptions that it makes. Namely, it assumes that guessing and item discrimination have no bearing on the outcome. Proponents of Rasch assume that guess and discrimination are accounted for by random noise and, as the noise is randomly distributed, it does not affect measurement. Rasch analysis accounts for anomalies such as guessing by providing misfit statistics which allow for items that misfit the Rasch model to be excluded from the test.

Importantly for test development, the Rasch model theoretically assumes unidimensionality in that all questions on the scale are measuring the same construct. As each question is plotted along this dimension in terms of difficulty, this also means that pupil ability can be pinpointed. It follows that the pupil would be able to answer all previous questions and none of the subsequent questions. This has real application in the development of tests to help children learn as teachers can identify what the pupils need to work on to progress. A traditional approach to testing would merely report a score or percentage.

Although Rasch measurement has only a small but enthusiastic following, the model has been applied to many studies, particularly in the areas of education and health (see Kingsbury et al., 2009 , Clements et al., 2008, Vasilyeva et al., 2009 for some current examples). The number of followers of the Rasch model is increasing and this burgeoning interest taken together with its capacity to deliver a more reliable and consistent measure makes it an attractive option at the outset for the development of a new test.

1.8   Adaptive test development

It is likely that children will have a wide range of abilities in facial expression decoding. To get an accurate measure of ability across the range, it would be necessary to have a test composed of very many items. If a four year old was to be asked 100 questions about shapes, you could be very confident that within that range there would be some questions that were easy for them and some that were difficult and so you would obtain an accurate picture of their ability. However, traditional paper-based methods of testing are limited in the number of items that can be included as it is unreasonable to expect a small child to attend to a test for more than a few minutes. It is more likely that a child is asked four or five questions on shapes.

If all four year olds' mathematics ability scores were plotted on a graph, you would expect to see a bell curve or 'normal' distribution. This would also be the case for other naturally

occurring phenomena such as height and weight. The diagram below shows the normal distribution. Pupils are represented by blue circles. The majority of children will fall in the 'average' range. A small number of children will fall in the 'low' or 'high' ability range.
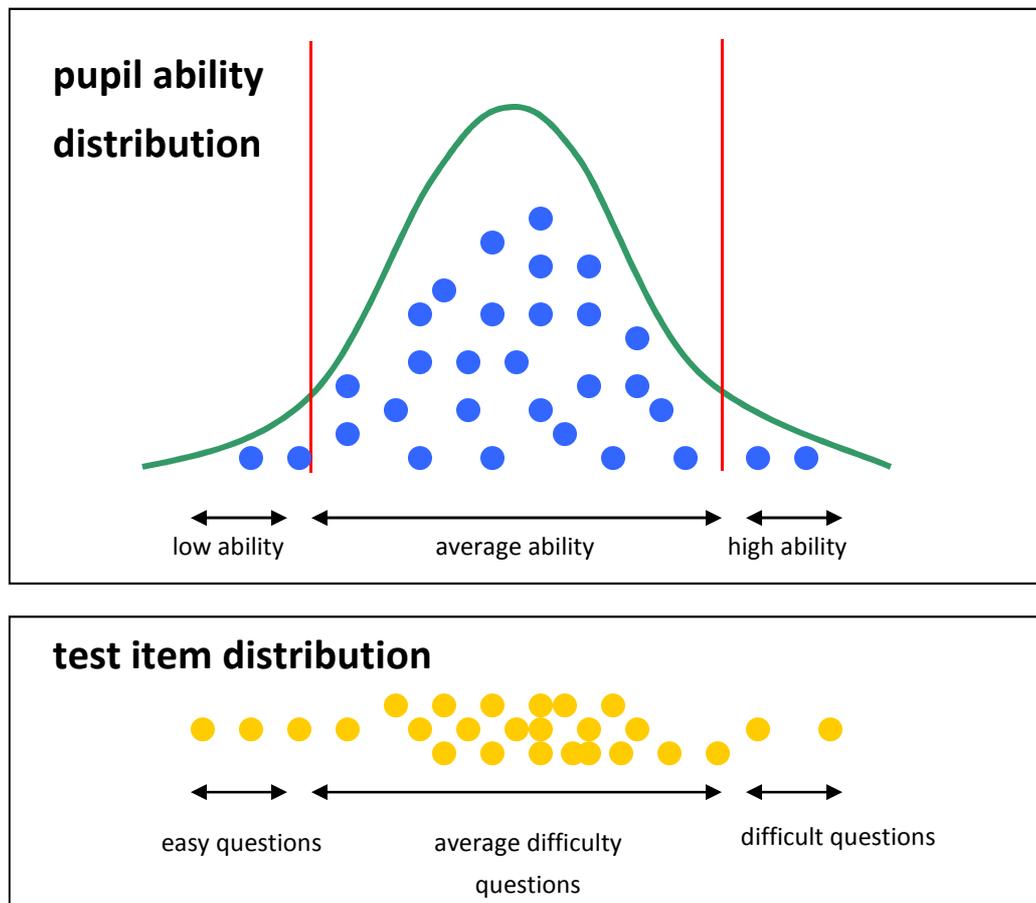
Figure 4: Representation of traditional test delivery

A traditional test would usually have a lot of test items that are suitable for the average ability children because there are so many more items. In Figure 4, each test question is represented by an orange circle. As the diagram demonstrates, there are substantially fewer children with low or high mathematics ability and only a few items will typically be included that are very difficult for the high ability children, or very easy, for the low ability group. In practice, this means that low ability children will be faced with a test that has very few items that they are able to answer which can be very damaging to their confidence and self-esteem. Gifted children, on the other hand, are not challenged by enough sufficiently demanding items. A further disadvantage is that having fewer items at the extremes of the ability range reduces the reliability of the test.

An alternative to the traditional approach to test delivery is to use adaptive testing. Adaptive tests aim to present all children with enough items of an appropriate difficulty to give an accurate measure of ability. A further advantage of this method is that because the items are more targeted, fewer items are needed which makes the test shorter. An adaptive test would be difficult to deliver in paper format without using a one-to-one

interview which can be complex and time consuming as the administrator needs to work with each test-taker, carefully monitoring and recording how each question is answered, then using this information to work out the next question to be presented.

Adaptive testing is ideally suited to computerised delivery. Tests can be created which apply pre-determined algorithms to select questions that depend on the individual's response to the previous question. Adaptive testing can be diagrammatically represented as shown in Figure 5.

pupil ability distribution

low ability   average ability   high ability

test item distribution

easy questions   average difficulty questions   difficult questions
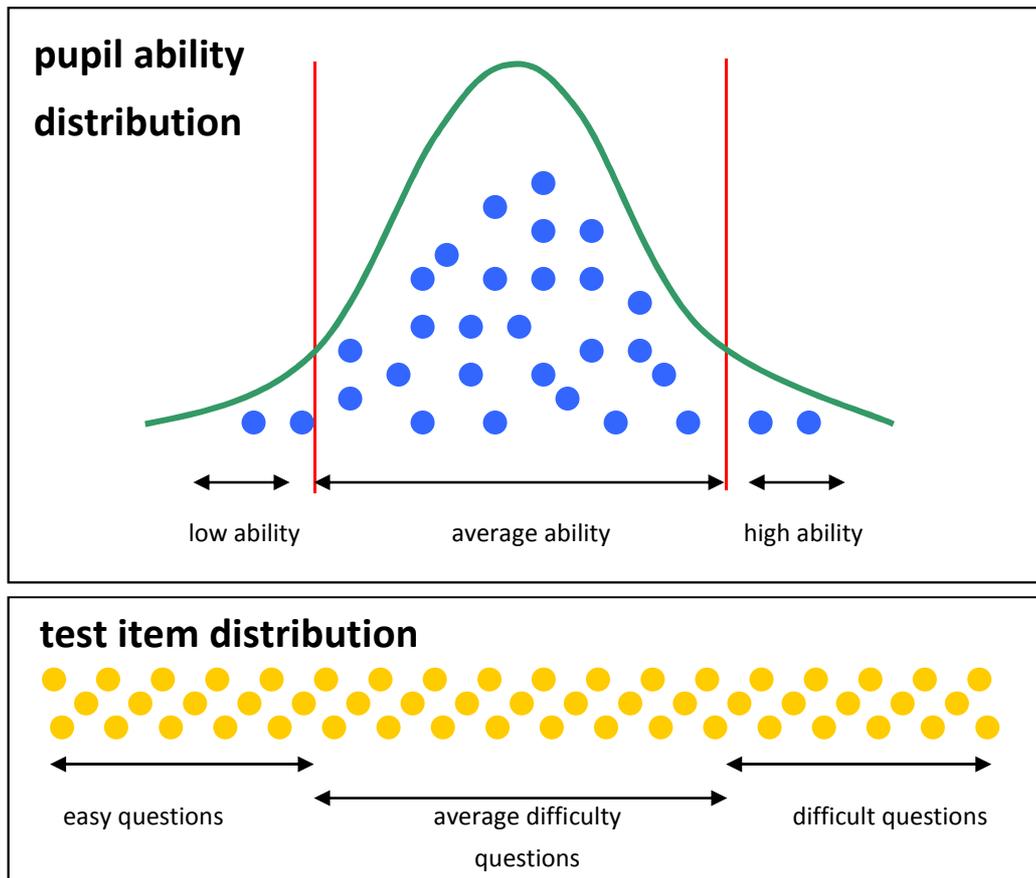
The test would consist of a bank of very many items covering a wide ability range and Rasch measurement, as described earlier, provides the necessary measurement framework. The facility values or measure of difficulty (as determined by Rasch) would be used to deliver the first question based on a pre-determined starting point such as a child's age. If the child answers correctly, a harder question would be retrieved from the hidden item bank and delivered on-screen. If the child answered incorrectly, an easier question would appear. This process would continue delivering items to target the child's ability and can do so using far fewer items than a traditional test. Research has found this method to be an innovative and efficient way to measure ability (Merrell and Tymms, 2007a).

1.9   Summary

During the process of the literature review, several strands of research were identified as being of particular interest.

Several studies have suggested that facial expression decoding is linked to emotional intelligence more widely.

There is little literature that adopts an approach to facial expressions that differs from the universality theory. The application of universality theory has proven successful in identifying correlates with certain behaviours and can be used to effectively discriminate ability both in people with disorders and those who are normally-developing.

Without exception, the studies fail to acknowledge the benefit that the involvement of individuals with the research process might have. The positivist stance regularly adopted has imposed adult assumptions on child participants. These studies often seek to generalise to a wider population but this should not be at the expense of understanding participants and acknowledging their position as individuals rather than subjects. It is possible that more reliable and valid data could be gathered through developing more appropriate instruments and this cannot be achieved effectively without the involvement of children.

There is little consideration given to playing to children's strengths. If assessment tools such as those described are to be commonly used in classrooms it is essential that they are appropriate and enjoyable. It is likely children will perform more reliably on a test that they can access and enjoy. This highlights a possible tension between the validity of the assessment and the accessibility by the pupils.

The use of language, controlled for with differing success in the studies, may have compromised the research. This is particularly relevant with young children whose language is developing and when comparing individuals from diverse cultural and language backgrounds.

The Rasch model presents an attractive alternative to traditional methods and computer-adaptive testing may be superior to traditional paper delivery.

1.9.1   Research questions

The literature review addressed relevant, current thinking in EI and test development to form the basis for this research. Firstly, the EI definition posited by Mayer and colleagues was accepted and, further, their specific ability model was identified as a sound basis for further research in the area. Ekman's basic expressions theory was found to prevail in the literature and be useful as a framework to measure facial expression decoding. It was decided that it was appropriate to measure individual pupil differences in this area with a view to contributing information to a pupil profile to help teachers identify those at risk. Finally, Rasch measurement, with its inherent assumptions of unidimensionality and the

relevance of pupil ability, was selected as an appropriate method for the development of a scale. From this starting point, several research questions were identified:

- Can a reliable and valid scale be developed to differentiate the ability of young children to correctly decode emotional facial expressions?
- Can this instrument be developed to minimise the confounding effects of language?
- Can the instrument be developed to be attractive and engaging for young children?

2 Development of the scale

2.1 Introduction

The focus of this research was to investigate whether a reliable and valid scale could be developed to differentiate children's ability to decode facial expressions of emotion. Review of the existing body of literature suggested that little research has investigated this ability in very young children without language playing a key part in the assessment and hence, acting as a confounding factor. It was also found that much work in the field has been dominated by a positivistic approach which has failed to consider young children's strengths. This section of the study will address several stages in the development of a culture- and language-reduced instrument suitable for use with young children. Central to the development of the instrument was a qualitative element to construction of the test which involved children in the preparation of items. This was incorporated to ensure that the instrument would expose true competencies rather than imposing an adult interpretation on how items should be developed and marked, which then coloured data analysis and subsequent interpretations.

2.2 Ethics

The research was conducted in line with guidance issued by the ESRC in the Research Ethics Framework (2007) and Durham University's Policy on Good Conduct in Research (2011). A proposal and Ethics Approval form were submitted to Durham University Ethics Committee and approval was received. A copy of the ethics proposal can be found in Appendix 5.1. The research involved working with children in a school environment in two ways. Interviews and group discussions were held with a view to developing items and visual stimuli. Children were also involved as participants in the trialling of the new test. Both stages required negotiating access to participants.

Schools were identified through their involvement with academic monitoring systems provided by the Centre for Evaluation and Monitoring (CEM) at Durham University. In the first instance, the head teachers of the participating schools were approached informally. After verbal agreement was received a formal letter outlining the research to be carried out was sent to the head teacher. A letter and an information sheet that outlined the purpose of the research and gave reassurances of anonymity and data protection were sent to the head teacher for distribution to the parents of the prospective participants. Copies of these letters can be found in Appendices 5.2 and 5.3. The documentation made clear that the study was entirely opt-in, with children, teachers and schools being able to withdraw at any time. It was also emphasised that data would be held securely in compliance with the Data Protection Act 1998 and that participants would be anonymised in any reporting of the research. Parents were assured that no judgments would be made about their children as a

result of the study and evidence of a criminal status check for the researcher would be made available.

The research involved visiting the participating schools and working with children in small groups and one-to-one. Before the interviews and discussions and data-collection phase of the study, the researcher spent some time familiarising herself with the staff and children with a view to minimising the fear of unfamiliarity and enable the children to become involved with the research without discomfort. However, during the entirety of the study, the researcher monitored the children for any signs of distress and looked to the classroom teacher (who was present or nearby throughout) for similar cues.

Although the consent of the head teacher and parents had been obtained, the children were made aware of the aims of the research. At the start of the visits, the researcher explained that she was hoping to make a new test and that she would like their help. The nature of the test was not revealed as it may have jeopardised the research process. She explained the research process, what was involved and reassured the children that they were able to withdraw at any time. The children were also reassured that no judgments would be made about them from the research findings. The assistance of the classroom teacher was sought to ensure effective communication with the children and to help identify any children who may have difficulties in understanding the process or be unduly worried by the researcher's presence.

The ideal situation for collecting data for this research was to gain access to children through local schools, firstly, because were the test to be successfully developed it would be intended for use by teachers in a classroom setting.

Secondly, being able to access large groups of children in a school setting was beneficial. Limitations of time and resources meant that fewer research sessions with larger groups was necessary and it was important to have enough participants to generate reliable data as the aim was to be able to generalise to a wider population. However, anyone conducting research in a school environment must be sensitive to issues of power. Children are expected to conform within the school environment although a researcher coming into the school cannot sidestep ethical procedures for testing by using the school as a proxy for access. Because children are expected to attend to their teachers and carry out instructions often without question, the children are likely to see the researcher as another teacher and behave similarly. Correct ethical practice, however, would determine that a researcher must negotiate the child's participation just as if that participant were an adult in their own home. The researcher was sensitive to this issue and was careful to treat each child as she would an adult participant.

Field notes were taken throughout the research and occasionally audio-recording equipment was used to assist with retrospective analysis of field notes. These notes and recordings were kept securely in line with the Data Protection Act and labelled with a code which did not allow for identification of individuals or schools.

2.3 Study One

2.3.1 Method

It was determined at the outset of this study that the instrument would be computer-delivered. There were several reasons for reaching this decision. As the aim of this study was to produce a useful measure for schools, it was sensible to consider the format in which this test would be made available for wider use. The increasing pervasiveness of computer technology in schools suggested that producing a paper-based test would immediately put the instrument at a disadvantage when compared to other measures. As described earlier, computer delivery, particularly computer-adaptive testing, offers an attractive alternative to paper based methods for reasons of measurement and efficiency. Computerised delivery minimizes the need for adult intervention; it has the advantage of helping to standardise administration of the test and can record the child's responses accurately and automatically.

The first stage of development, therefore, included a small scale initial trial of the new test with a view to identifying whether the content and computer-delivered format of the assessment were appropriate for the age of the children.

2.3.1.1 Instrument

In order to ensure that the effects of culture and language were minimised, it was considered crucial that items were developed with this in mind while harnessing children's true ability as far as possible. An innovative paradigm was employed using cartoon vignettes.

Cartoons were selected in preference to photographs for a number of reasons. Firstly, it was considered important that the questions were culturally appropriate. Using cartoons made it possible to limit cultural specificity of the characters. Secondly, children are accustomed to reading books with cartoon drawings in them and seeing cartoon characters on the television and it was reasoned that the format of the images would be familiar to them and attractive enough to keep their attention. The use of cartoons also enabled tweaks to be made to the stimulus material based on field testing. Finally, the cartoons facilitated the design of the innovative test delivery method which is described in further detail below.

Twenty emotional vignettes were drawn depicting cartoon characters in situations that suggested the character was experiencing one of the six basic emotions as defined by Ekman. The character depicted in each vignette had no facial detail. The vignettes can be found in Appendix 5.5.

A set of 13 cartoon facial expressions was prepared. The style of face was based on the work of Herman Chernoff (1973) who used simple facial representations to represent multivariate data. His reasoning was that humans are able to identify facial variations expertly and so small changes in data would be detected, when represented by, say, a

widening of smile. The faces modelled by Chernoff were found to be interpreted consistently and were thought to provide a good basis for development.

One cartoon was created for each of the six basic emotions plus a neutral expression. The literature had suggested that children may be able to differentiate between differing levels of expression intensity. Therefore, in order to discriminate ability further, two levels of intensity for each facial expression were created, one being more subtle than the 'standard' expression. The expression cartoons can be found in Appendix 5.6. The faces were trialled with ten adults, five male and five female, to check for validity. Participants were shown the cartoon expression and asked to name the emotion they saw in the expression. There was a high level of agreement between the participants for all of the expressions although the naming of the 'disgust' face caused some difficulty with some participants mimicking the expression and being able to say what made them feel like that but only agreeing on disgust when prompted. It is possible that, rather than a difficulty with the concept of the item, this reflected the naming of the word. A word frequency database was checked to compare the frequency of the emotion labels. The results are shown in the Table 5 and suggest that, indeed, the difficulty with the disgust label may be attributed to word frequency.

Table 5: Word frequency of emotion labels

Emotion label	Word frequency (frequency per million)
Happy	419
Sad	238
Fear	51
Surprise	230
Angry	130
Disgust	3

The items for the test were created by showing a selection of four of the cartoon expressions alongside each of the twenty vignettes. One of the faces had an expression most suitable for the emotion depicted in the vignette. The other three expressions were taken from the remaining bank of emotional expressions and the neutral face. Each vignette appeared with a standard target among standard distracters, with the standard target among subtle distracters, with a subtle target among standard distracters and with subtle target among subtle distracters. Difficulty of the items was estimated during development with a view to being able to discriminate a range of abilities. For example, an easy 'sad' item showed a vignette where the cartoon character had dropped his ice-cream. The multiple choice face options were sad, happy, very happy and disgust. A difficult question showed a cartoon character hiding under a table during a thunderstorm. The options were sad, angry, fear and neutral. There was more similarity between the facial expressions for this question than for the easy question. Some example items are shown in Figures 6 and 7.
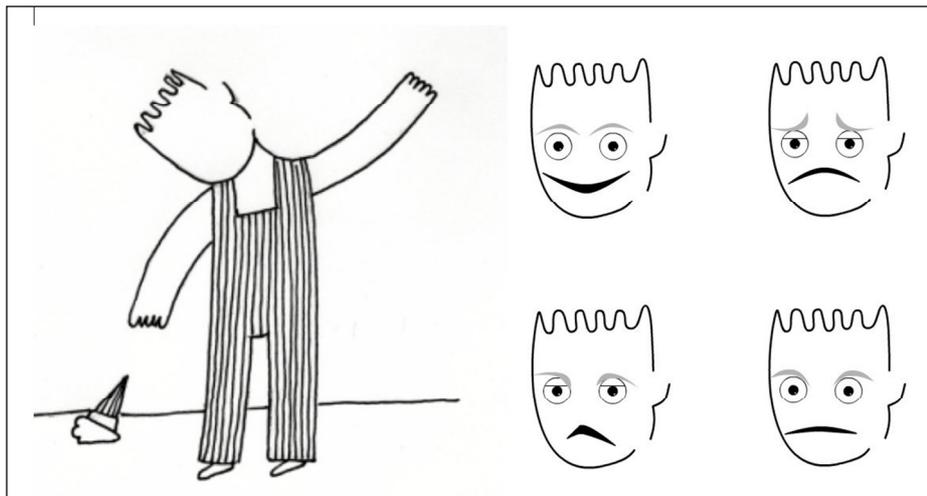
Figure 6: Example 'sad' item for FACES 1.0
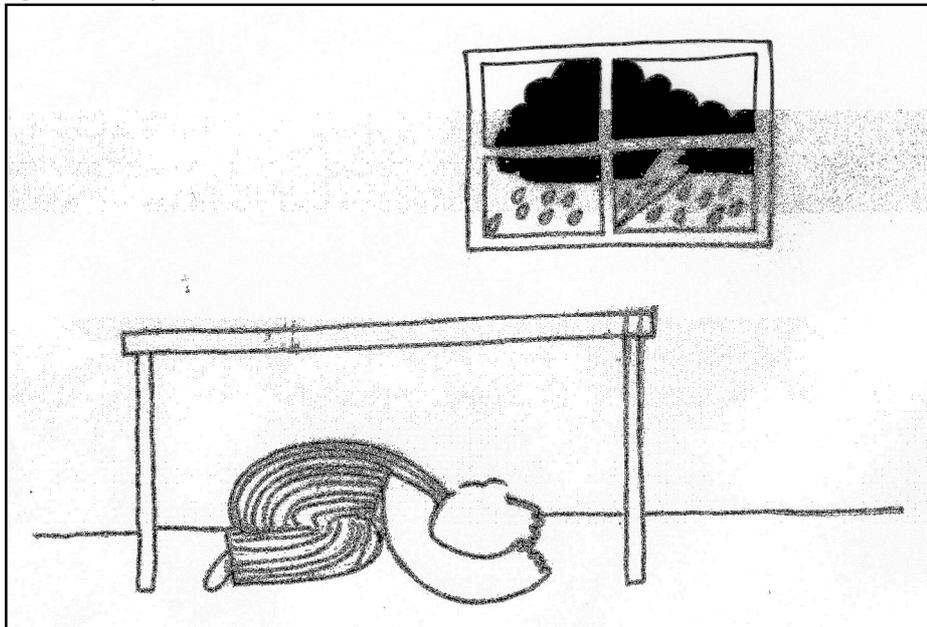
Figure 7: Example 'fear' item for FACES 1.0

The test was delivered using a piece of software under development called CADATS (Computer Assisted Design and Testing System) (He, 2006) which was specifically created to help teachers with the design and delivery of computer-based tests. Each of the items was put together as an individual Flash[2] file. These files were then inserted into CADATS. At the outset of each test delivery session, the software presented a screen used to collect pupil identifier, date of birth, sex and school and then moved onto the test. The items were presented in a random order to control for fatigue.

---

[2] Flash is a file format used to deliver video or picture clips and which can be embedded within computer programs

31

Initial pilot

A small-scale pilot of the 78 items was conducted in one school with 23 children aged 4 and 5 years. As the children were being guided through the assessment, they were encouraged to talk through their decisions and any difficulties they were having which provided some very useful initial information on the items. Several items proved to be too difficult for the children and, although the pilot test had good item reliability (Cronbach α = 0.88), some of the items had poor discrimination. Poorly discriminating items were adjusted by changing the selection of expressions. Table 6 shows the breakdown of items in the revised bank.

Table 6: Number of items from each emotion included in FACES 1.0

Emotion	Number of items
Happy	12
Sad	19
Angry	16
Fear	13
Disgust	9
Surprise	9
Total	**78**

The construction of the first four items is exemplified in Table 7. The full item list can be found in Appendix 5.4 and the corresponding artwork can be found in Appendices 5.5 and 5.6.

Table 7: Item bank for FACES 1.0

Item no	Picture	Target	Distracter 1	Distracter 2	Distracter 3
1	sad1	sad	happy	angry	disgust
2	angry1	angry	sad	neutral	happy
3	disgust1	disgust	angry	neutral	happy
4	fear1	fear	disgust	happy	neutral

It was clear from the initial pilot that a 78 item test was unmanageable for the small children. Although they enjoyed the experience initially, they soon tired and became distracted. It was necessary to reduce the number of items in the test although this presented a challenge. If the test were to be made available as a computer-adaptive test in the future, it would be necessary to gather information such as item correlations, discrimination measures and facility values on a large number of questions to create a bank of items to form the basis of the adaptive test engine. Therefore, in order to reduce the assessment period to approximately 10 – 15 minutes, and still gather data on all the items, the test was redesigned to incorporate a set of anchor items presented to all children with a random selection of items presented from the remaining bank. The anchor test comprised 28 items which were judged to cover a range of difficulty. They were weighted in the following way: two items portraying happy scenarios, nine sad, six angry, three fear, four disgust and four surprise. The program then presented one of four sub-sets, each

comprising 10 items which included at least one item for each of the six emotions. The items were presented in a random order, but in the same order for each participants. These adjustments from the initial pilot resulted in the FACES 1.0 test, delivered using the CADATS test-delivery software.

2.3.2 Participants

Convenience sampling was again employed to gain access to local schools. Again, schools were identified through their involvement with academic monitoring systems provided by the Centre for Evaluation and Monitoring (CEM) at Durham University.

The participants were 61 children in four different schools. In order to minimize disruption in the schools, the head teachers were asked whether a sample of children could be tested rather than the whole class. The class teachers were asked to make the selection to cover a wide range of academic ability. Although the test was being designed with a view to it being used in the Reception class (with 4 to 5 year olds) participants from Year 1 (age 5 to 6) were included in the sample to identify whether the test could be used with older children too. Details of the sample are outlined in Table 8.

Table 8: Participants in the trial of FACES 1.0

School	Reception		Year 1		Total
	Girls	Boys	Girls	Boys	
School A	15	12	8	6	41
School B	3	1	1	3	8
School C	1	1	4	5	11
School D	0	0	1	0	1
Total					**61**

In order to see whether the participants were typical for their age in terms of ability and achievement, two different assessments were compared against nationally representative norms. The two assessments were the PIPS On-entry Baseline (PIPS BLA), conducted at the start of the Reception year, and the PIPS End of Year 1 Assessment (PIPS Y1).

The PIPS BLA (Merrell and Tymms, 2007b) is run by the Centre for Evaluation and Monitoring at Durham University. The assessment is conducted in the first six weeks of children beginning full-time, compulsory education. It is an attractive computer-delivered assessment which is administered on a one-to-one basis. The assessment contains items that have been found to be good predictors of later success or difficulty at school and reflect the general developmental level of the children. The following areas are assessed:

- Writing – the child is asked to write his/her own name and the quality of the writing is scored against examples.
- Vocabulary – the child is asked to identify objects embedded within the picture.
- Ideas about reading – assesses concepts about print.
- Repeating words – the child hears a word and is asked to repeat it in this assessment of phonological awareness.

- Rhyming words – the child selects a word to rhyme with a target word from a choice of three options in this assessment of phonological awareness.
- Letter identification – a fixed order of mixed upper and lower case letters.
- Word recognition and reading – this starts with word recognition and moves on to simple sentences that the child is asked to read aloud. The words within these sentences are high frequency and common to most reading schemes. This is followed by a more difficult comprehension exercise which requires the child to read a passage and at certain points select one word from a choice of three that best fits that position in the sentence.
- Ideas about mathematics – assessment of understanding of the vocabulary associated with mathematical concepts.
- Counting and numerosity – the child is asked to count four objects. These disappear from the screen and then the child is asked how many objects they saw. This is repeated with seven objects.
- Sums – addition and subtraction problems presented without symbols.
- Shape identification.
- Digit identification – single, two digits and three digits.
- Maths problems – including sums with symbols.

The PIPS Year 1 assessment is presented either on paper or computer. It is carried out within a two week window in June and is administered either on a class or group basis. The assessment measures mathematics and reading using items that are based on the English National Curriculum. A measure of developed ability is also included. This contains items for capturing vocabulary and non-verbal ability. Together, these measures claim, with some justification, to reflect the child's capacity to learn, distinguished from academic achievement (Tymms, 2002). The scores of the children in this sample are reported in Table 9. PIPS scores are standardised on nationally representative samples of children completing the same assessment at the same time of year. The mean score of the national sample is 50 and the standard deviation is 10.

Table 9: Mean achievement and ability scores of the participants in Trial One

Assessment	Study sample mean	standard deviation
PIPS On-entry Baseline total score	56.0	9.9
PIPS On-entry Baseline Follow-up total score	66.5	7.1
PIPS End of Year 1 Mathematics	53.2	8.8
PIPS End of Year 1 Reading	57.1	10.4
PIPS End of Year 1 English vocabulary	55.5	9.8
PIPS End of Year 1 Non-verbal ability	52.7	7.9

The mean total score of the children in the Reception group for this study was higher than the national average at the start of the year. During the Reception year those particular

children moved further ahead compared with the national sample and their mean score was one and a half standard deviations higher than the national average. The children in Year 1 were also slightly higher than average for mathematics and non-verbal ability.

2.3.3 Procedure

The FACES 1.0 test was given to 61 children in four schools over the course of three days. Testing was conducted in November to ensure that pupils were relatively settled into school and comfortable in their surroundings. Individual testing sessions were scheduled in consultation with the class teacher to ensure minimum disruption to the school day. Testing was either carried out in a quiet corner of the classroom or in another space agreed with the teacher and within sight of a member of staff. The test was delivered to children individually using a laptop computer operated by the researcher. Before each test was administered, the researcher explained the aims of the research using age appropriate language and ensuring the child understood their right to withdraw at any stage. The researcher monitored the child for signs of anxiety or fatigue throughout the process. Each testing session lasted approximately 10 minutes.

At the end of each testing session, the researcher talked with the child to find out whether they enjoyed the test, if it worried them, whether they found the format easy to understand and whether they liked the pictures and computer delivery.

The researcher made field notes where possible throughout the testing sessions and conducted informal interviews with the children.

2.3.4 Findings

2.3.4.1 Data from trial of FACES 1.0

Item response data was gathered from the 61 participants. Each item was coded to represent the choice of face made by each participant. The responses were then marked 1 for correct and 0 for incorrect. WINSTEPS software (Linacre, 2011) was used to apply the Rasch model to the data.

Rasch measurement analyses internal reliabilities differently from the traditionally reported Cronbach alpha (α) which cannot be used where individuals answer different sets of questions. The Rasch model measures internal reliability in two ways. The first indicates the level of confidence that participants who score highly on the test have a high ability and participants with low scores have low ability. This is referred to as the person reliability measure. It also shows whether items of high difficulty are consistency difficult and that low difficulty items are consistently easy. This is referred to as item reliability. The internal reliabilities of the FACES 1.0 scale were 0.40 (person) and 0.57 (item).

In order to be able to contrast the performance of the FACES 1.0 scale with other measures, the 28 anchor items were analysed separately in SPSS and found to have internal reliability of 0.54 (Cronbach α).

The item level results for the full FACES 1.0 test (anchor test and sub-tests) are shown in Table 10. Relative difficulties of each item are reported as a percentage of participants who were presented with the item and answered it correctly. The higher the percentage, the easier the item. The lower the percentage, the more difficult the item. Discrimination values are also reported. In order to fit the Rasch model, WINSTEPS assumes that all item discriminations are equal with a value of 1. Empirically, however, item discriminations vary and WINSTEPS reports an estimate of those discrimination values which makes it possible to identify items that do not fit the Rasch model. The further the value away from 1 (either above or below) the less discriminating the item is. A value over 1 suggests that the item is discriminating between high and low ability participants more than expected for a question of that level of difficulty. If the value is below 1, the item is discriminating less than expected for a question of that level of difficulty. The correlation measure represents how well each item in turn correlates with others in the scale. Items with a correlation of less than 0.2 may not correlate well with the other items. However, there are difficulties with eliminating items purely on this basis because item correlations are affected by how many times the item has been used in the test and the distribution of ability of participants.

Table 10: Item facility and discrimination and correlation values for FACES 1.0

Item reference code	Number of times item presented	Presentation type (A=anchor, 1,2,3,4=subset)	Facility %	Discrimination	Correlation
Q1Sad	61	A	36	0.74	0.25
Q2Angry	62	A	27	1.23	0.25
Q3Disgust	62	A	35	0.8	0.24
Q4Fear	62	A	79	1	0.15
Q5Happy	62	A	18	0.99	0.17
Q6Surprise	62	A	73	0.94	0.21
Q7Sad	62	A	16	1.02	0.21
Q8Angry	62	A	27	0.98	0.23
Q9Disgust	62	A	42	1.12	0.28
Q10Fear	62	A	68	0.96	0.27
Q11Happy	10	4	100	1.33	0.34
Q12Surprise	61	A	43	1.05	0.18
Q13Sad	62	A	32	1.11	0.26
Q14Angry	60	A	45	0.72	0.25
Q15Angry	61	A	52	0.99	0.23
Q16Happy	22	2	0	1.11	0.13
Q17Sad	61	A	30	1.39	0.22
Q18Sad	61	A	56	1.01	0.19
Q19Angry	61	A	41	0.88	0.28
Q20Disgust	61	A	34	0.19	0.23
Q21Fear	23	1	74	1.03	0.2
Q22Happy	23	1	74	1.15	0.2
Q23Surprise	61	A	5	1	0.13
Q24Sad	61	A	49	0.97	0.18
Q25Angry	22	1	41	0.79	0.27
Q26Disgust	9	4	44	0.88	0.36
Q27Fear	11	5	73	1.03	0.15
Q28Happy	11	5	91	1.12	0.14
Q29Surprise	51	A	8	1.06	0.2
Q30Sad	51	A	53	1.08	0.19
Q31Angry	13	1	23	1.33	0.15
Q32Fear	12	3	67	1.04	0.22
Q33Happy	12	3	75	1.03	0.16
Q34Sad	49	A	47	0.99	0.25
Q35Sad	13	1	31	2.04	0.17
Q36Angry	13	1	46	1.03	0.16
Q37Disgust	13	1	23	1.07	0.18
Q38Fear	12	2	83	1	0.17
Q39Happy	12	2	8	1.02	0.13
Q40Surprise	13	1	77	0.99	0.11
Q41Sad	-1	4	0	1	
Q42Angry	12	3	8	1.22	0.18
Q43Disgust	12	3	42	1.12	0.25
Q44Fear	13	1	46	1.04	0.13
Q45Happy	13	1	92	0.95	0.1

Q46Surprise	12	3	25	1.3	0.24
Q47Sad	12	2	25	1.2	0.2
Q48Angry	10	4	70	1.02	0.23
Q49Fear	10	4	50	1.29	0.21
Q50Happy	10	4	70	1.32	0.18
Q51Sad	62	A	23	1.48	0.23
Q52Sad	12	2	83	1.02	0.17
Q53Angry	12	3	58	1.25	0.18
Q54Disgust	12	2	42	1.23	0.2
Q55Fear	12	3	92	1.05	0.15
Q56Happy	12	3	17	1	0.22
Q57Surprise	12	2	100	0.99	0.15
Q58Sad	11	5	64	0.88	0.27
Q59Angry	11	5	27	1.25	0.18
Q60Disgust	11	5	45	0.73	0.21
Q61Fear	12	2	75	1.08	0.22
Q62Happy	61	A	84	0.97	0.11
Q63Surprise	10	4	0	0.97	0.15
Q64Sad	12	3	83	1.02	0.17
Q65Angry	61	A	62	0.92	0.22
Q66Fear	11	5	73	1.01	0.15
Q67Happy	11	5	73	0.93	0.19
Q68Sad	10	4	50	0.41	0.32
Q69Disgust	62	A	42	0.91	0.21
Q70Angry	62	A	2	1.04	0.19
Q71Angry	12	2	33	1.36	0.22
Q72Sad	12	3	58	0.93	0.18
Q73Sad	11	5	55	0.96	0.24
Q74Angry	10	4	50	-0.31	0.3
Q75Fear	10	4	30	1.18	0.3
Q76Sad	13	1	15	1.22	0.15
Q77Angry	11	5	9	0.86	0.29
Q78Surprise	11	5	45	1.46	0.15

Examination of facility, discrimination and correlation information suggested that the items shown in Table 11 should be removed.

Table 11: Items to be removed from FACES 1.0 following examination of facility, discrimination and correlation data

Item reference code	Number of times item presented	Presentation type (A=anchor, 1,2,3,4=subset)	Facility %	Discrimination	Correlation
Q16Happy	22	2	0	1.11	0.13
Q20Disgust	61	A	34	0.19	0.23
Q23Surprise	61	A	5	1	0.13
Q28Happy	11	5	91	1.12	0.14
Q35Sad	13	1	31	2.04	0.17
Q39Happy	12	2	8	1.02	0.13
Q40Surprise	13	1	77	0.99	0.11

Q44Fear	13	1	46	1.04	0.13
Q45Happy	13	1	92	0.95	0.1
Q62Happy	61	A	84	0.97	0.11
Q68Sad	10	4	50	0.41	0.32
Q74Angry	10	4	50	-0.31	0.3

Analysis of the revised test with WINSTEPS produced internal reliabilities of (0.49) person and 0.66 (item). Removing the poorly discriminating and correlating items had improved the internal reliability of the scale, but the reliabilities were still low.

The misfitting items in the scale were then examined. These are items which do not conform to the Rasch model and therefore may affect the performance of the test. The analysis addresses misfitting items at two levels. The outfit statistic is sensitive to extreme items (for example, very difficult or very easy items) while the infit statistic is influenced by the pattern of responses to each item. If badly fitting items are removed, internal reliability of the test should be improved.

Items that have an outfit statistic of below 0.8 or above 1.3 can affect the performance of the measurement. Seven items had outfit statistics of below 0.8. However, several of these items had only been seen by around 10 participants and it was decided to remove only two items that fell below 0.7.

Wright offers a useful way of interpreting infit statistics by categorising items into four bands which describe the effect that those items would have on the measurement (Linacre, 2011). Wright's guidelines are shown in Table 12.

Fit statistics	Effect on measurement
>2.0	Distorts or degrades the measurement system.
1.5 – 2.0	Unproductive for construction of measurement, but not degrading.
0.5 – 1.5	Productive for measurement.
<0.5	Less productive for measurement, but not degrading. May produce misleading good reliabilities and separations.

The infit statistics from this analysis were applied to Wright's model. The findings were encouraging and suggested that only one further item be removed. This item had previously been identified as outfitting.

The items that were removed are shown in Table 13.

Table 13: Items to be removed from FACES 1.0 following examination of misfit statistics

Item reference code	Number of times item presented	Presentation type (A=anchor, 1,2,3,4=subset)	Infit	Outfit
Q11Happy	10	4	0.38	0.13
Q22Happy	23	1	0.76	0.73
Q42Angry	12	3	0.81	0.79
Q46Surprise	12	3	0.71	0.68
Q50Happy	10	4	0.7	0.71
Q53Angry	12	3	0.87	0.73
Q61Fear	12	2	0.87	0.79

The data was once again analysed using WINSTEPS which gave internal reliabilities of the FACES 1.0 scale to be 0.37 (person) and 0.62 (item). This suggested, rather than improve the scale, item and person reliability had been further degraded.

An item-map was produced through WINSTEPS and is shown in Figure 8. The item-map plots relative difficulty of the items against the relative ability of the pupils. The items are displayed on the right hand side of the scale and the distribution of pupil ability on the left. The higher the logit value, the more difficult the item or the higher the ability of the pupil. The letter 'M' denotes the mean, 'S' is one standard deviation from the mean and 'T' is two standard deviations from the mean.

The anchor items have been highlighted as they were completed by all the children and so will be less susceptible to the effect of sample size.

```
           PERSON - MAP - ITEM
              <more>|<rare>
     1              +
                   |T Q70ANGRY
                   |  Q33HAPPY  Q57SURPR
                   |  Q21FEAR   Q29SURPR   Q48ANGRY   Q56HAPPY   Q65ANGRY
                   |  Q6SURPRI  Q8ANGRY
              .  T|S Q36ANGRY  Q52SAD     Q78SURPR
             ###   |  Q32FEAR   Q43DISGU   Q49FEAR   Q54DISGU   Q73SAD
           .#### S|  Q24SAD    Q31ANGRY   Q3DISGUS  Q51SAD    Q58SAD
                   |  Q72SAD    Q76SAD
     0    .#####   +M Q12SURPR  Q15ANGRY   Q34SAD     Q47SAD    Q55FEAR
                   |  Q69DISGU  Q7SAD
      .########## M|  Q17SAD    Q18SAD     Q19ANGRY  Q26DISGU   Q27FEAR
                   |  Q2ANGRY   Q66FEAR    Q9DISGUS
          .#### S|  Q13SAD    Q1SAD      Q37DISGU  Q4FEAR    Q60DISGU
                   |  Q64SAD    Q77ANGRY
          .#  |S Q14ANGRY  Q25ANGRY   Q67HAPPY  Q71ANGRY
          #  T|  Q10FEAR   Q59ANGRY   Q75FEAR
                   |  Q30SAD    Q5HAPPY
                   |T
    -1         .  +  Q38FEAR
                   |  Q63SURPR
                   |
                   |
                   |
                   |
                   |
    -2              +
              <less>|<frequ>

Key: anchor item
```

Figure 8: Item-map from Trial One of FACES 1.0

The item-map shows that the distribution of item difficulties was appropriate for the distribution of pupil ability. This can be reasoned as mean difficulty is aligned with mean ability and there is a range of anchor items to cover the range of ability.

The literature review identified that skill at decoding the six different expressions was acquired at different rates. The item-map is repeated in Figure 9 with the different emotions colour coded.
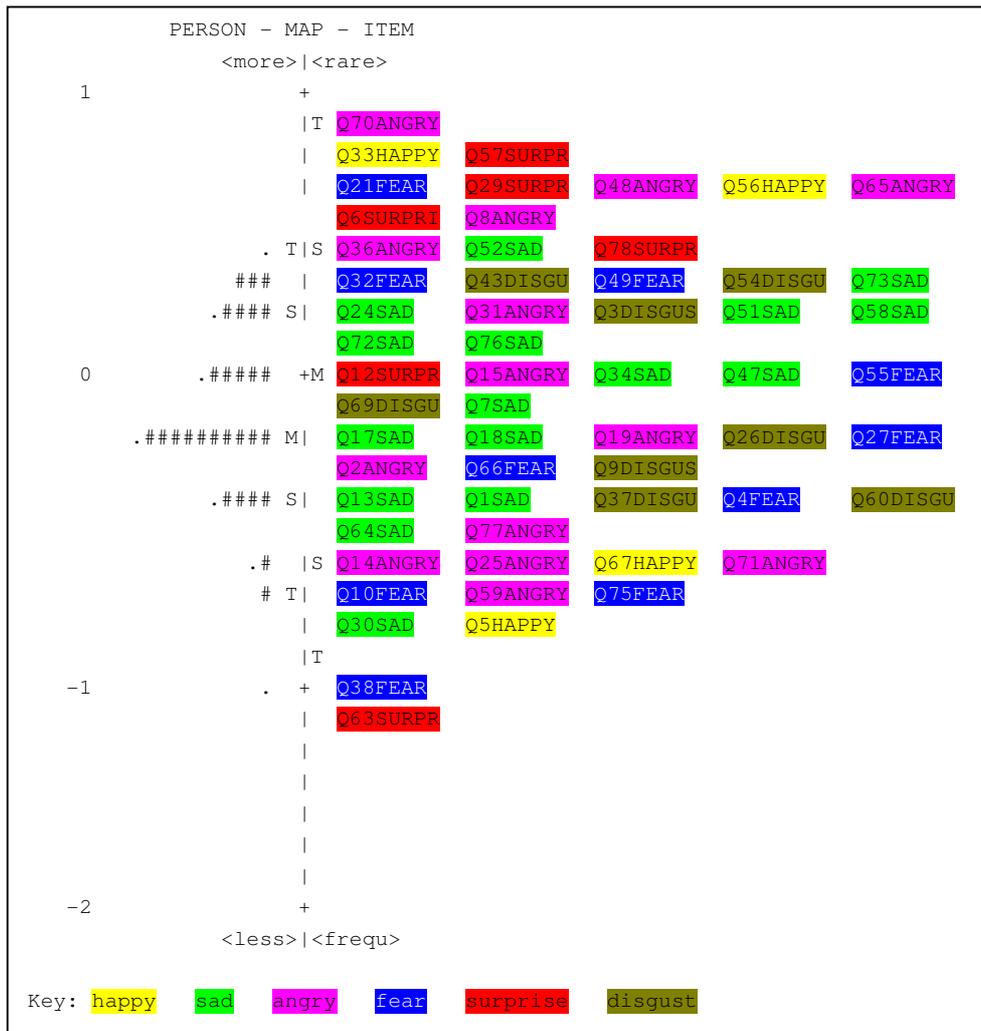
Figure 9: Item-map from Trial One of FACES 1.0 with colour coding of emotions

The colour coded item-map does not suggest that any of the facial expressions were more difficult to decode than others. It should be noted that the sample was small for some of these items and that further data on all items would be useful for interpreting these findings.

Fundamental to the usefulness of Rasch for building a measurement instrument is the assumption that the scale is unidimensional, i.e. it is measuring one factor. In this case, it is necessary to determine that the measurement is of ability to decode all facial expressions in others (the first factor) and not, for example, a second factor of ability to decode expressions of anger. WINSTEPS uses principal components analysis (PCA) to identify whether there are clusters of items that appear to be addressing a second factor. This analysis is sensitive to sample size. A rule of thumb for use of PCA is given here:

As the sample size for this trial was 61, PCA was considered an appropriate measurement to detect second or subsequent factors in the instrument.

PCA applies measures of variance to identify additional dimensions (or factors) in the data and these findings are reported as shown in Table 14. The first column shows each stage of the variance analysis. The second column reports the findings in Eigen value units. This unit describes the strength of the secondary dimension and roughly maps onto number of items. Therefore an Eigen value of 2.0 suggests a secondary dimension of two items. Interpretation of this value would depend on the number of items in the test. The third column reports percentage of variance explained. The higher the percentage, the more variance explained by the dimension.

Table 14: Principal components analysis of FACES 1.0

Stage of analysis	Variance in Eigen value units	Variance as %
Raw variance explained by measure	12.6	17.8
Raw variance explained by persons	4.5	6.4
Raw variance explained by items	8.0	11.4
Raw unexplained variance	58.0	82.2
Unexplained variance in 1st contrast	4.3	6.2
Unexplained variance in 2nd contrast	3.9	5.6
Unexplained variance in 3rd contrast	3.2	4.5
Unexplained variance in 4th contrast	2.8	1.0
Unexplained variance in 5th contrast	2.8	4.0

PCA of FACES 1.0 raised some concerns. Only 17.8% of the variance in the data could be explained by the measure. The Eigen value unit suggested that only 12 of the items in the test were measuring the first dimension (which was assumed to be the ability to decode all facial expressions). This suggested the possibility of more dimensions in the test. It may have been, for example, that decoding expressions of fear was tapping into a different set of processes.

PCA involves several iterations of a process that examines the data for additional dimensions and this is reported in Table 14 as 'unexplained variance in contrast'. Each level of contrast represents a further dimension in the test. The findings above show the contrasts had strengths of four Eigen value units or less, which in a 78 item test does not suggest significant additional dimensions. Where additional dimensions are not found, the Rasch model predicts that unexplained variance can be accounted for by random noise. This has implications for the test construction as it suggests that, although over 80% of the differences between children's scores were not explained by the test, this unexplained

variance was distributed equally over the sample and allowed for appropriate interpretation of the results.

2.3.4.2   Observation of testing

Although the instructions for the assessment were carefully explained to the children, they were monitored throughout the process to check that they had remembered the procedure. All children appeared to be quite happy with the process of matching one of the faces to the vignette although some appeared to find it more challenging than others and took considerably longer.

Several children appeared to be puzzled when the same vignette appeared more than once. The structure of the test meant that although the items were all unique, the vignettes were repeated with different distracters. When the same vignette appeared for a second time, many of the children commented that they had already done that question when, in fact, it was a new item. The test presented the items randomly which meant that sometimes the same vignette appeared twice in a row. Sometimes the children thought that the computer had stuck on the same question. The tests took around ten minutes per child. Many of the children were seen to become distracted and began to fidget and shuffle about towards the end of the test.

2.3.4.3   Discussions with participants

During the course of the testing, the children were questioned briefly and informally about some of their responses to the items with the intention of checking for validity. The children were, on the whole, quite happy to explain their choice which highlighted some difficulties with the vignettes. There was some confusion between the 'fear' and 'surprise' vignettes. Vignette 'fear 2' shows the character sitting at a table with a spider appearing above him. Some of the boys said that they would find this exciting so selected the happy response rather than the allocated answer, 'fear'. However, one particularly perceptive little girl said that "*well I think spiders are nice, but other people would think they are frightened so I chose that face*".

Some children were able to offer a correct verbal label for the expressions that they chose although there was some variation. For example, 'cross', 'fed up' and 'annoyed' were used to label the angry expression. When asked about the disgust questions, the children struggled to offer any label although this applied equally to children who had selected the correct response and those who had selected a distracter. The vignette for 'disgust 2' appeared to be problematic with several children unable to explain what they thought was happening in the picture.

2.3.4.4   Post-testing interviews

At the end of each testing session, the researcher conducted brief informal interviews with each child to establish whether they enjoyed the test and gather their views on the format, pictures and computerised delivery.

None of the children interviewed expressed any concerns or worries about the test. In fact, the majority of children reported that they had enjoyed the experience. One child noted that "*the man is funny, I like him*". Another said "*I like playing games like this, I do them at home*". Over half the children said that they enjoyed spending time with the researcher and liked being able to help.

The children were specifically asked about the computerised delivery of the test. All the children reported that they were comfortable with doing the test on the computer. Several of the children were very enthusiastic about computers and wanted to talk about things they had done in their lessons. One child said that she "*loves doing computer work, it is my favourite*". Another said "*I like doing things on the computer, can I do it again*?"

When asked about the cartoons used in the test, there were mixed reports. Most children were ambivalent about the pictures. Several children thought that the character was Bart Simpson "*but not yellow, he isn't very good is he*?" Around ten of the girls said that he was a little bit scary and one said "*I don't think the faces are very nice they are creepy*". One child said "*they would be nicer if there was some colour*".

2.3.5 Discussion of Study One

The intention of this study was to find out if a reliable and valid scale could be developed to differentiate the ability of young children to correctly decode emotional facial expressions. Further research questions were whether the instrument could be developed to minimise the confounding effects of language and could it be developed to be attractive and engaging for children. Trials were conducted in four schools and with 61 participants. The findings for each research question are discussed here.

2.3.5.1   Can a reliable and valid scale be developed to differentiate the ability of young children to correctly decode emotional facial expressions?

The findings were subjected to scrutiny based on the criteria suggested by Mayer *et al* (2008).

Systematic test design was addressed as set out in Table 3 earlier. Content validity was considered to be good. The FACES 1.0 scale was based on the body of theory on facial expression decoding as described by Ekman and the universals approach. The scale reflected the basic emotions that theory posited and incorporated different degrees of intensity. Evidence of response-process validity was gleaned through informal questioning of participants during the process of testing. It was apparent that the children understood the task being asked of them and were able to select a response that was appropriate to their consideration of the item. Some children were not able to offer a response to some of the questions and could not explain why but this would be expected in a test that was designed to separate out ability on a particular dimension. There were occasions where children's interpretation of the vignettes differed or their answer reflected different responses to a situation. For example, boys often provided an 'angry' response to a

situation that girls responded to with 'sad'. This could represent a function of differentiating ability but may also suggest further development of the vignettes with more input from children.

Examination of facility, discrimination, correlation and misfit allowed the scale to be refined to give internal reliabilities of 0.37 (person) and 0.62 (item), derived from Rasch measurement. The DANVA, LEAS-C and EMT discussed earlier reported reliabilities of between 0.64 and 0.93 (Cronbach α) with similar aged participants. WINSTEPS person reliability is equivalent to the Cronbach reliability measure reported in the other studies and shows the FACES 1.0 test to be considerably less reliable than the other instruments. However, the person reliability measure is sensitive to sample size and could have been affected by the number of responses to the items in the subsets. Running a further trial where all items are presented to all participants may improve the test properties although that would need to be balanced with a test of appropriate length for young children.

Rasch analysis of internal reliabilities is not exactly comparable with the Cronbach alpha. Rasch approximates reliability using standard error and is often an underestimation. Cronbach alpha employs analysis of variance and is known to overestimate reliability. Cronbach alpha could not be performed on the FACES 1.0 dataset because not all of the participants had seen all the items. However, the anchor items were separated out and found to have a Cronbach alpha reliability estimation of 0.55 which was more encouraging. The FACES 1.0 test was most similar to one subtest of the Emotion Matching Task, differing only in that verbal prompts were used at the presentation of each item. The Cronbach α for that subtest was 0.54.

Item reliability is sensitive to the length of the test and the sample size which make it difficult to place the items on a single scale. It is possible to improve item reliability by testing more participants and to increase the length of the test, but lengthening the test would be problematic as the current 38-item test was already proving a little too long for some of the small children.

Principal components analysis caused some concern with only 17% of the variance explained by the FACES 1.0 scale. The Rasch model would assert that the unexplained variance can be accounted for by random noise and that this is evenly distributed throughout the test so allowing for appropriate interpretation. It was not possible to evaluate this against the other measures addressed as none reported level of variance explained.

In their study, Mayer *et al* (2008) suggested that measures of emotional intelligence should incorporate factors and subscales if they are to represent EI as a true intelligence. The FACES 1.0 was found to have one factor which supports the supposition that decoding of facial expressions is a discrete skill. If this is the case, the FACES 1.0 test would be represented as a specific ability model in the review by Mayer and colleagues. It is also possible that it could contribute to measurement in an integrative model. The FACES 1.0

test might address emotion perception skills in the hierarchy of EI as set out by Mayer and colleagues in their model.

It was not possible to gather evidence of convergent validity for two reasons. Firstly, ethical approval was not given for administration of additional EI measures and secondly, financial and time constraints prevented the use of a second instrument.

The results from the Trial One suggest that a reliable and valid scale had not been achieved. However, comparing the results with other instruments measuring the same concept was encouraging.

2.3.5.2   Can the instrument be developed to minimise the confounding effects of language?

Once the procedure had been explained to the children, apart from the informal questioning, no language was used for the purposes of conducting the assessment. Three of the children assessed were second language English speakers and appeared to have no difficulties accessing the test. They attended to instructions appropriately, followed the procedure without further prompting and gave appropriate responses. As noted previously, the children had difficulties in labelling the disgust expression, but as some of the children provided correct responses by matching the faces, this language difficulty did not appear to affect their ability to answer the questions. This suggested that implementation of a language-reduced design was encouraging.

Of the existing instruments reviewed, the DANVA had the least reliance on language but the items were still dependent on accurate verbal labelling. The results of this study identified one child who could correctly match an expression of disgust to a visual vignette but was unable to offer a verbal label when questioned. The other tests would have marked this response as incorrect where the FACES 1.0 test allowed the child to respond without the confound of language ability. The DANVA, LEAS-C and EMT instruments do not claim to be testing EI separately from language skills. Indeed, parts of their models deal specifically with emotional language. However, it could be argued that their measures of EI are too highly dependent on language to access a pure cognitive ability particularly in young children or non-English speaking populations.

The EMT has a sub-test which presents a series of verbal vignettes and asks the child to provide a verbal emotion label. Although the authors claim that this is measuring an aspect of EI, it may be mediated by other cognitive processes. For example, the child may need to make use of their short term memory to recall elements in the vignette. The task in the EMT that is most similar to FACES 1.0 still requires the use of language in that the child is asked to provide a verbal label for a visual stimulus. In real life, children do not need to put a verbal label on their interpretation of a facial expression. Taken together, these difficulties suggest that other instruments relying on language are not providing an ecologically valid test. FACES 1.0 exploits naturally-occurring cognitive tasks. As the children are not being constantly prompted and are free to interpret the tasks at their own pace

using processes that are not too contrived, this may contribute to the low level of variance explained in the analysis.

The results from Trial One were encouraging and suggested that a language-reduced paradigm for assessing facial expression decoding was achievable.

2.3.5.3 Can the instrument be developed to be attractive and engaging for young children?

Computer delivery of the assessment was useful in standardising delivery and the children appeared to enjoy the process and, in general, engaged with the test. It was apparent, however, that the artwork was not entirely appropriate. In discussion, none of the children volunteered an active dislike of the pictures, but evidence from the post-testing interviews suggested that they could be improved and may be partly responsible for the children losing interest towards the end of the testing session.

Some children were puzzled by repetition of artwork but it is not clear whether this may have affected the performance of the test. Future versions of the test would need to remove any duplication of items or artwork. However, the CADATS software used to develop the test had limitations in that incorporating more items or colour images would slow down the delivery of the test to an unacceptable level.

The first trial suggested that although the children happily took part in the assessment and were comfortable with the paradigm, improvements could be made to the stimuli to accommodate their tastes and preferences.

2.3.6 Conclusions

Several conclusions were drawn as a result of Study One which provided ways forward for further development.

- Although reliability measures were low, they were similar to those found in other sub-tests claiming to measure the same cognitive ability but which relied more heavily on the use of language.
- To give more confidence in interpreting the properties of the test, a larger sample size would be useful. Ideally, the test would present all items to all participants.
- A measure of convergent validity would be needed.
- The content of the test would benefit from more input from children to strengthen the validity.
- The test interface could benefit from improvements to make it more appealing and engaging for young children.

2.4 Study Two

The next stage of the research involved taking forward the developments suggested in 2.3.6 and running a further trial. This process is further described here.

2.4.1 Method

The findings from Trial One fed into three areas of focus for Trial Two. Firstly, it was suggested that the properties of the FACES 1.0 scale were encouraging although reliability measures were too low for the results to be interpreted and used with confidence. The findings from Study One suggested that some improvements were needed to the items to ensure the suitability of the content for young children. It was anticipated that improved items trialled with a larger sample size would have a positive impact on the reliability of the instrument. It was also posited that further involvement of children in the development of items was required as a second area of focus. One of Mayer's criteria for evaluating instruments was to consider a measure of convergent validity which was not possible as part of Study One. This was the third area of focus for Study Two.

2.4.1.1 Instrument

The paradigm used in FACES 1.0 was maintained but some developments were implemented.

The cartoon vignettes were reconstructed with two key considerations. Firstly, children were involved in the adaptation of existing items and the development of new items. Secondly the cartoon vignettes and faces were redrawn in colour and with a different style to be more attractive for the young children.

Initially, six new cartoon expressions were drawn by a graphic artist under the guidance of the researcher. In order to remain as culturally unbiased as possible, the cartoon face was coloured in a neutral tone. Each of the six expressions, plus a neutral control, was printed out onto A4 paper. The cartoons can be found in Appendix 5.8.

The researcher worked with a group of 28 children in one local school. Year 1 children were chosen as they were close to the target age group for the FACES test but had been in school for a year and were more settled and happier to communicate with the researcher. The full class was involved in an attempt to cover a range of ability. The children were asked if they would like to participate and all agreed. The children were split into four groups to make the discussion more manageable and gave more opportunity for every child to voice his or her opinion. The researcher worked with each group in turn on their circle time carpet. The children were shown the series of six cartoon faces in turn. With each expression, the children were asked what might be happening to the character to make them feel that way. Verbal labels for the expressions were not used by the researcher at any point during the session. The children were encouraged to give the first response, rather than the researcher giving an example that might influence the children's interpretation of the expression. Once one child had offered a response, plenty of other suggestions were offered. The researched noted all of the responses until repetition and no further ideas suggested saturation.

Overall the children's responses reflected the emotion that was appropriate for the vignette. Importantly, this validated the new facial expressions and also generated a list of

possible scenarios to use as a basis for the development of existing vignettes and the creation of new ones.

Eighteen new cartoon vignettes were drawn, three vignettes for each of the six basic emotions, some adapted from existing items. Each cartoon was presented with a set of four cartoon faces. One of the faces had an expression that was most appropriate for the emotion depicted in the vignette. The three distracters were taken from the remaining five expressions and the neutral expression. As in Trial One, two versions of each facial expression were used, one being subtle and one standard. Each vignette appeared with a standard target among standard distracters, with a subtle target among subtle distracters, with a subtle target among standard distracters and with a standard target among subtle distracters. This resulted in a bank of 72 items, 12 for each of the 6 emotions. An example item is shown below.

Figure 10: Example 'disgust' item used to construct FACES 2.0

During the creation of the new cartoons, a selection of the images was shown to a small sample of six children ranging in ages between four and seven (known personally by the researcher). The children were interviewed individually. They all reacted positively to the pictures by asking questions about what was happening or smiling and when asked if they liked them, they all replied that they did. Although it was a small group, involving the children during the development was very helpful as they were able to point out where elements of drawings were ambiguous and needed improvement.

As was highlighted in Study One, a 72 item test was too long for young children. Even the 38 item test was appearing to be too long for some children. It was decided to limit the test to 36 items. The intention at the outset of this research was to develop an adaptive test. To do this, a large number of items are needed and these must be trialled with a large number of children to determine the test properties. It was decided that, to gather enough data for

meaningful analysis, at least 100 children should be tested across four schools to give a reasonably representative spread of ability and home background. The test designed for Trial One included 28 anchor items delivered to all children and then one of four subsets consisting of 10 items from the remaining item bank. In order to get enough data on the subset items, 400 children would need to be tested. Within the limitations of time and resources available for this thesis, testing to this extent was not feasible. A decision was made for Trial Two to use a flat test design, as opposed to adaptive, to gather sufficient data on all items to be able to determine the potential of the test.

Six items were chosen to represent a range of difficulty within each emotion category. Table 15 shows the breakdown of items in the new item bank.

Table 15: Number of items from each emotion included in FACES 2.0

Emotion	Number of Items
Happy	6
Sad	6
Angry	6
Fear	6
Disgust	6
Surprise	6
Total	36

The construction of the first four items for the FACES 2.0 test is exemplified in Table 16. The full item list can be found in Appendix 5.7 and the corresponding artwork can be found in Appendices 5.8 and 5.9.

Table 16: Item bank for FACES 2.0

Item no	Picture	Target	Distractor1	Distractor2	Distractor3
1	angry1	angry	sad	neutral	happy
2	angry1	very angry	sad	neutral	happy
3	angry2	angry	neutral	fear	happy
4	angry2	very angry	neutral	fear	happy

The CADATS software used for the delivery of the FACES 1.0 test in Trial One included many features not relevant and was undergoing continual development which required considerable support. This would have made it unfeasible as a product for use in the classroom. A decision was made to simplify the data collection and analysis process by delivering the new version, the FACES 2.0 test, through a tailored interface which was programmed using Visual Basic language. This made the test more robust, and easier to run and install.

An introductory screen was generated to collect biographical information on each pupil. The test began immediately following a page of instructions intended to guide the adult operating the computer. Images of these two screens can be found in Appendix 5.10. All 36 items were presented to all participants in random order. The Visual Basic programme

speeded up the presentation of the items which meant the whole test took around six minutes per child.

A simple scale for teachers was developed to provide a measure of convergent validity. The regular classroom teacher was asked to rate children on their ability to identify the emotions of others on the basis of their observations during the school year. They were given a pupil list and asked to score 1 for the five children most able, 2 for those not identified and 3 for the five children least able. The reasoning here was that teachers would quickly be able to identify children at the extremes. To ask the teachers to rate each individual on a scale was considered too onerous a task and unlikely to provide an accurate picture. The data collection form used can be found in Appendix 5.11.

2.4.1.2 Participants

Convenience sampling allowed four local schools to be accessed for Trial Two. None of the four schools had been involved in Trial One. Researchers assessed as many children as possible in one day beginning with the Reception class and moving on to Year 1 if time permitted. The Reception class was tested first as the test would be intended to identify children in need of help as early as possible. Some Year 1 pupils were also assessed to gauge the suitability of the test with older children. Details of the sample are given in Table 17.

Table 17: Participants in the trial of FACES 2.0

School	Reception		Year 1		Total
	Girls	Boys	Girls	Boys	
School A	9	8	11	10	38
School B	13	22	13	12	60
School C	9	23	2	2	36
School D	16	20	0	0	36
Total					**170**

In the first trial, PIPS test data were used to look at whether the ability of the pupils was typical for schools in the UK. This ensured that the development of the test was being guided by a representative sample from the population. This was not possible in the second trial as the schools were not involved in the PIPS Project. Data was obtained from the Department of Education for the purposes of establishing representativeness of the sample. The data used were aggregate test percentages across the three core subjects (English, maths and science) that reported how many pupils achieved the expected level or above at Key Stage 2 during 2007. Although this data did not relate to the same pupils as were used in this study, it was considered to act as an appropriate proxy. The data are shown in Table 18.

Table 18: No of pupils achieving expected level or above at Key Stage 2 during 2007

Sample	aggregate % of pupils achieving expected level or above max = 300%
England average	245
School A	225
School B	300
School C	176
School D	198

(Source: Department for Education, Department for Education, 2007)

The data is limited in that it does not report subjects separately and an aggregate percentage is not a convenient statistic for comparison purposes. However, it does suggest a range of ability across the four schools. Pupils in School B outperformed the national average, with all pupils achieving the expected level or above. School C appeared to have considerably less able pupils than average.

2.4.1.3 Procedure

Two researchers conducted the testing following the procedure as outlined in 2.3.3. During the testing, the researchers observed the children and took notes where appropriate. The observations were carried out by the author and another researcher not involved in the development of this instrument. At the end of each period of testing, the researchers compared notes to gain an overall picture of the assessment process.

2.4.2 Findings

2.4.2.1 Data from Trial Two

Item response data was gathered from 170 participants and the internal reliabilities of the FACES 2.0 scale were 0.75 (person) and 0.91 (item). Cronbach α was 0.77.

Item level results for the full FACES 2.0 test are shown in Table 19.

Table 19: Item facility, discrimination and correlation values for FACES 2.0

Item reference code	Number of times item presented	Facility %	Discrimination	Correlation	Infit	Outfit
Q1Angry	170	57	1.15	0.34	0.96	0.97
Q2Angry	170	61	0.88	0.34	1.04	1.02
Q3Happy	170	66	1.13	0.33	0.96	0.91
Q4Happy	170	69	1.26	0.32	0.89	0.83
Q5Happy	170	70	1.4	0.32	0.82	0.75
Q6Happy	170	70	1.3	0.32	0.87	0.80
Q7Happy	170	54	1.13	0.34	0.97	0.96
Q8Happy	170	60	1.04	0.34	0.98	1.00
Q9Sad	170	44	0.9	0.33	1.03	1.01

Q10Sad	170	34	0.89	0.32	1.04	1.05
Q11Sad	170	32	0.56	0.31	1.16	1.24
Q12Sad	170	26	0.91	0.29	1.04	1.09
Q13Sad	170	51	1.04	0.34	1	0.98
Q14Sad	170	41	0.9	0.33	1.02	1.05
Q15Surprise	170	46	1.28	0.34	0.94	0.92
Q16Surprise	170	41	1.29	0.33	0.92	0.92
Q17Surprise	170	28	0.56	0.3	1.2	1.36
Q18Surprise	170	26	0.55	0.3	1.19	1.49
Q19Surprise	170	41	0.88	0.33	1.03	1.05
Q20Surprise	170	45	1.33	0.34	0.92	0.92
Q21Angry	170	46	0.75	0.34	1.05	1.08
Q22Angry	170	35	0.93	0.32	1.02	1.03
Q23Angry	170	53	0.89	0.34	1.02	1.04
Q24Angry	170	56	0.95	0.34	1.01	1.01
Q25Disgust	170	39	0.89	0.33	1.04	1.00
Q26Disgust	170	54	1.12	0.34	0.97	0.97
Q27Disgust	170	43	0.82	0.33	1.04	1.07
Q28Disgust	170	44	0.77	0.33	1.05	1.06
Q29Disgust	170	39	1.03	0.33	0.99	0.98
Q30Disgust	170	48	1.35	0.34	0.92	0.92
Q31Fear	170	55	1.6	0.34	0.86	0.84
Q32Fear	170	56	1.16	0.34	0.96	0.97
Q33Fear	170	55	1.13	0.34	0.97	0.96
Q34Fear	170	52	1.17	0.34	0.96	0.96
Q35Fear	170	30	0.88	0.31	1.04	1.10
Q36Fear	170	34	0.91	0.32	1.03	1.04

Examination of this data suggested that no items needed to be removed. The facility values showed that the items covered a range of difficulty although there appeared to be no very easy or very hard questions. The items were correlated at an acceptable level with no item having a correlation of less than 0.2. The discrimination values were encouraging with the items clustering quite closely around 1 and suggesting that the items fit the Rasch model. No outfitting items were identified and Wright's interpretation of infit statistics suggested that all items were productive for measurement.

An item-map was produced and is shown in Figure 11:

```
            PERSON - MAP - ITEM
               <more>|<rare>
   2             +
                 |
                 |
            X    |
                 |
                 |
            X    |
                T|
            X    |
                 |
           XXX   |T
                 |  Q12SAD
   1       XXXXX +  Q18SURPR
          XXXXXXXXX |  Q17SURPR
                 |  Q35FEAR
            XXXX |
                 |  Q11SAD
         XXXXXXXX S|S Q10SAD    Q22ANGRY  Q36FEAR
        XXXXXXXXXXX |
                 |  Q29DISGU
       XXXXXXXXXXX  |  Q16SURPR  Q25DISGU
          XXXXXXX  |  Q14SAD     Q19SURPR
                 |  Q27DISGU  Q28DISGU  Q9SAD
         XXXXXXXXXX |  Q15SURPR  Q20SURPR  Q21ANGRY
   0     XXXXXXXXXX +M
         XXXXXXXXXX |  Q30DISGU
               M|  Q13SAD
      XXXXXXXXXXXXX |  Q23ANGRY  Q34FEAR
         XXXXXXXX  |  Q26DISGU  Q31FEAR    Q33FEAR    Q7HAPPY
                 |  Q1ANGRY    Q24ANGRY  Q32FEAR
           XXXXXX |
       XXXXXXXXXXX |S Q8HAPPY
                 |  Q2ANGRY
           XXXXXX |
                 |
        XXXXXXXXX S|  Q3HAPPY
  -1       XXXXXX +
                 |  Q4HAPPY   Q5HAPPY    Q6HAPPY
           XXXXXX |T
                 |
          XXXXXXX |
                 |
            XXX   |
                 |
                T|
            XX   |
                 |
            X    |
  -2             +
                 |
            X    |
                 |
                 |
                 |
                 |
                 |
                 |
                 |
                 |
                 |
  -3             +
              <less>|<frequ>
```

Figure 11: Item-map from trial of FACES 2.0

The item map showed the distribution of item difficulties to be appropriate for the range of pupil ability with the mean difficulty and mean ability aligned. There was a range of items to cover the ability range although the item map would appear to confirm that there are no

items that are appropriate for the very high or low ability children. However, the test appeared to have identified a group of around 14 pupils with low scores which was the intention of the measure.

The item-map is repeated in Figure 12 with colour coding to identify whether the FACES 2.0 test supported the theory that skill in facial expression decoding was acquired at different rates for the different emotions. The item-map would suggest that, overall, there was a spread of emotion items across the scale, although children appeared to find the happy faces easier to decode and the surprise faces slightly harder.

Figure 12: Item-map from trial of FACES 2.0 with colour coding of emotions

Principal components analysis was conducted and the results are shown in Table 20.

Table 20: Principal components analysis of FACES 2.0

Stage of analysis	Variance in Eigen value units	Variance as %
Raw variance explained by measure	7.2	16.6
Raw variance explained by persons	2.3	5.3
Raw variance explained by items	4.9	11.3
Raw unexplained variance	36.0	83.4
Unexplained variance in 1st contrast	2.0	4.6
Unexplained variance in 2nd contrast	1.9	4.4
Unexplained variance in 3rd contrast	1.8	4.2
Unexplained variance in 4th contrast	1.7	4.0
Unexplained variance in 5th contrast	1.7	3.9

As in Trial One, the percentage of variance explained was low at 16.6%. The analysis would suggest that seven of the items appeared to measure the ability to decode facial expressions. There would not appear to be further dimensions present which suggest that the unexplained variance could be attributed to random noise.

Convergent validity was measured using the teacher rating scale. The FACES 2.0 test scores of the children who were rated most able by their teachers were compared against those rated least able using an independent t-test of the total raw score. The results are shown in Table 21:

Table 21: Independent T-test comparing least able and most able pupils on a teacher rating scale

Group	No of pupils	Mean total score	Standard deviation	Standard error mean
Least able	19	14.05	6.06	1.39
Most able	18	20.33	5.10	1.20

$p = .002$

There was a significant difference between the groups of r= 6.28, which was a large effect size (1.12). This suggested that the teacher ratings were separating out ability as the pupils they rated least able were getting lower scores on the test and higher rated pupils were getting higher scores.

2.4.2.2 Observations of testing

The researchers noted that the children quickly grasped the task of matching the face to the situation and appeared to enjoy taking part. A few children appeared to lose concentration towards the end of the assessment but the majority of children maintained an appropriate level of attention throughout.

Although the children were not formally questioned during the assessment, some children did verbalise their train of thought which gave an interesting insight. As with the first trial, it appeared that some children were interpreting the pictures in different ways. In one instance the 'Angry 2' vignette (a cartoon of a girl scribbling on the other's painting) was interpreted as sad by one of the pupils. The disgust face was interpreted by two children as the character crying or being in pain.

The random presentation of the items meant that some of the vignettes were presented twice in succession. On one occasion a vignette appeared three times in succession. Some of the children found this frustrating because they thought it was a repeat of the same question. Despite this, in the cases observed, the children provided appropriate answers. Although some of the children were very talkative during the testing, after the initial explanation, none of the children needed further help with how to answer the questions. Some of the children were inquisitive, asked questions about the characters and wanted to tell stories around the pictures. One little girl asked why the target character was a boy and not a girl.

The test took around 5 or 6 minutes to complete with each child.

2.4.3 Discussion of Study Two

The intention of the first trial was to analyse the test properties of the FACES 1.0 instrument to determine its suitability as a reliable and valid scale to differentiate the ability of young children to correctly decode emotional facial expressions. The findings suggested that the test required revision to make it appropriate. Although the language-reduced paradigm was successful, improvements needed to be made to improve the test properties of FACES 1.0 and make the test more attractive for young children.

Trial Two was conducted after improvements had been made to the stimuli and the delivery of the test. The trial was conducted with 170 pupils in four schools. The findings are discussed here.

2.4.3.1 Can a reliable and valid scale be developed to differentiate the ability of young children to correctly decode emotional facial expressions?

Mayer's criteria for judging reliability and validity (as shown earlier in Table 3) were used to evaluate the FACES 2.0 test.

The new version of the scale was rooted in the same body of theory as the original FACES 1.0 test as described in 2.3.5.1. The assessment was considered to reflect the theory of six basic emotions as posited by Ekman which verified the content validity.

Response process validity was judged through observation of the children and confirmed the findings of Trial One that the children understood the task being asked of them and were able to select a response that was appropriate to their consideration of the item. Again, some children were not able to offer a response to some of the questions but a range of ability in this skill would be expected. Although the children were not formally questioned during the testing, observation raised similar issues to those identified in the first trial. Some pupils were interpreting the vignettes differently from how they were intended and some were puzzled by the repeated presentation of the same item. There were occasions where children's interpretation of the vignettes differed or their answer reflected different responses to a situation. For example, boys often provided an 'angry' response to a situation that girls responded to with 'sad'. This could represent a function of

differentiating ability but may also suggest further development of the vignettes with more input from children.

A few children appeared to lose concentration towards the end of the assessment. It would be expected that children's level of attention would differ with maturation and as the majority of children maintained an appropriate level of attention throughout, the scale was judged to be a suitable length.

Examination of facility, discrimination, correlation and misfit of the improved scale gave internal reliabilities of 0.75 (person) and 0.91 (item), derived from Rasch measurement, and Cronbach α of 0.77. This was a considerable improvement on the FACES 1.0 test. Indeed the FACES 2.0 test appeared to have test properties in line with those reported by the DANVA, MSCEIT and EMT instruments which reported reliabilities of between 0.64 and 0.93 (Cronbach α) with similar aged participants. The emotion matching subtask of the EMT was the most similar to the FACES 2.0 test and reported Cronbach α of 0.54. This task relied more heavily on the use of language and was less reliable than the FACES 2.0 test. It was assumed that the improved items and delivery would improve the test properties but delivering all items to a larger sample is likely to have made a contribution.

Principal components analysis again showed there to be a low level of variance (16.6%) explained by the FACES 2.0 test. The Rasch model would propose random noise to account for the unexplained variance. It was not possible to evaluate this against the other measures addressed as none reported level of variance explained.

2.4.3.2 Can this instrument be developed to minimise the confounding effects of language?

The findings from the second trial supported and extended those of the first; that the language-free paradigm developed was accessible to young children. The second trial involved a larger sample and once the task had been described to the children, they appeared able to progress through the test without further instruction.

2.4.3.3 Can the instrument be developed to be attractive and engaging for young children?

The children appeared to enjoy using the FACES 2.0 test. Children were monitored throughout and none appeared worried, distracted or unhappy. They smiled and asked questions. Most children maintained an appropriate level of engagement throughout, with only a few children losing concentration towards the end of the test.

There remained some issues with the stimuli. Vignettes were included more than once in the test construction and, because of the random presentation; they sometimes appeared twice in succession although it was a different item with different distracters. This puzzled a few of the children although they did appear to be able to answer the question without difficulty.

During the first trial, observations were carried out by the author. It is possible that this coloured the observations and findings from discussions. For the second trial, the author and a researcher not involved in the FACES development carried out the testing and observation. It was encouraging that the second researcher noted a similar level of enjoyment and engagement by the children.

3 Discussion

This research has reported on two studies concerned with the development of FACES, an innovative test to discriminate the ability of young children to decode facial expressions of emotion. The aim of the test was to identify deficits in facial expression decoding to allow for remediation at an early stage to improve outcomes for children.

The research centred around three questions:

- Can a reliable and valid scale be developed to differentiate the ability of young children to correctly decode emotional facial expressions?
- Can this instrument be developed to minimise the confounding effects of language?
- Can the instrument be developed to be attractive and engaging for young children?

3.1 Can a reliable and valid scale be developed to differentiate the ability of young children to correctly decode emotional facial expressions?

3.1.1 Reliability

A set of criteria was established for evaluating the test properties of the FACES instrument against existing measures, the framework based on criteria suggested by Mayer and colleagues.

Trial One found the FACES 1.0 test to have internal reliabilities of 0.37 (person) and 0.62 (item), derived from Rasch measurement. Person reliability can be aligned with more traditional measures of internal consistency such as Cronbach α. Other instruments claiming to measure EI in children have reported Cronbach α of 0.64 - 0.71 (LEAS-C), 0.88 (DANVA) and 0.88 (EMT). When anchor items were separated out the scale was found to have a Cronbach α of 0.55. This was still weak when compared to the other instruments and was a cause of concern, however, the other measures were mediated by language at different levels. The EMT included a subtask that was similar to the FACES paradigm. This involved matching pictures, and language was not used other than to give a verbal prompt at the presentation of each new item. A comparison of the internal consistency of this subtask (0.54) and the FACES 1.0 test (0.55) was more encouraging.

Refinement of the items and adjustments to the delivery of the test resulted in FACES 2.0. The reliability of the new version was found to be considerably improved. The internal reliabilities of FACES 2.0 derived from Rasch measurement were 0.75 (person) and 0.91 (item). The scale was now separating out ability to 2 or 3 levels (on a scale of 1 to 4). Cronbach α of 0.77 was reported, which gave FACES 2.0 a higher reliability that the most similar subtask taken from the existing, established measures of EI (the emotion matching subtask of the EMT). The MSCEIT is arguably the most widely used measure of EI and, although test properties for the MSCEIT-YV (Youth Version) have not yet been published, a

task in the adult version involving perception in faces has a higher language load with a Cronbach α of 0.80. The FACES 2.0 test is as reliable as the most similar task claiming to measure the same concept. The DANVA, LEAS-C and EMT involved a considerably more complex administration with several sub-tasks comprising the overall measure. The FACES 2.0 scale was providing a reliable measure with only a 5 or 6 minute test.

The improvement in the reliability of the scale after the refinements to version 1.0 were impressive. There were several possible explanations for this. Firstly, the items themselves had been improved. During Trial One, it became clear that improvements could be made to make the pictures more appealing and engaging for the age group, and the cartoons redrawn to make them more friendly and appealing with more accessible expressions. It is possible that this was rewarded by an increased level of attendance to the test. Secondly, the artwork was validated by the children. The expressions themselves were verified by the children and were used as a prompt to suggest content for the vignettes to match those expressions. This ensured that the children were able to respond to the vignettes in a way that was real to them and was not coloured by the interpretation of the researcher. Finally, all children were presented with all items. Although this meant that an adaptive testing approach could not be implemented at this stage, it is likely that it contributed to the higher reliability.

3.1.2 Validity

Validity was evaluated using Mayer and colleagues' criteria.

3.1.2.1 Content validity

Content validity was considered in order to determine whether the FACES scale was accurately reflecting the body of theory around EI. Results from Trial One suggested that content validity was good, with FACES 1.0 representing the current thinking in emotional intelligence, the universals approach and particularly the conceptual framework suggested by Mayer and Salovey. FACES 2.0, although improved from the first version, did not differ in its representation of theory, and content validity was considered to be appropriate. Indeed, it may have added to the body of knowledge in that stimuli were developed with the help of children who are the real specialists in interpretation of emotions in children. These findings would benefit from an independent assessment of content validity by other researchers in the field.

Mayer and colleagues suggested three models of EI measure; specific ability, integrative and mixed model. The FACES 2.0 test would fit, alongside the DANVA and LEAS-C, within the specific ability group of models which seek to identify discrete skills as components of EI and, arguably, measures the skill more effectively. It is equally possible that FACES 2.0 reflects a subtask within an integrative model such as the EKT or MSCEIT, although it may not be appropriate to make direct comparisons in terms of test properties as the integrative models are claiming to measure more than discrete skills. Assuming FACES 2.0 is

placed within the specific ability category, it is possible that facial expression decoding as represented by the FACES 2.0 test would have research application outside the EI framework and into other theories of emotion and cognition. This would extend its application outside the classroom and into further academic research.

3.1.2.2 Response-process validity

During Trial One, response-process evidence of validity was gained through informal questioning during the testing process. It was apparent that, overall, the children understood the task being asked of them and were able to select a response that was appropriate to their consideration of the item. Trial One did identify some items where response-process validity was less apparent: the children were struggling to make an appropriate interpretation. Improvements were made to the items for FACES version 2.0 in preparation for the second trial. Although the children were not formally questioned during the second trial, many of them did talk to themselves or to the researcher and these verbalisations suggested that response-process validity had been achieved.
The research suggested that the children were able to make an appropriate interpretation of the vignettes and were able to choose a response which corresponded with their judgment of correctness. The stimuli in FACES 2.0 were developed in collaboration with the children and it was hoped that this would have added to the response-process validity. However, for the purposes of this study, correctness was judged by the researcher. Validity could be verified further by comparing this with correctness as judged by the consensus scoring method which would ensure the children's views of correctness were taken into account.

3.1.2.3 Convergent validity

A weakness of Trial One was the lack of convergent validity. It was not possible at that time, due to constraints of time and resource, to deliver a further test to the children. Additionally, ethical approval was not granted for the use of an external instrument. For Trial Two a simple teacher rating scale was implemented which suggested that FACES 2.0 was accurately identifying the most and least able pupils. This finding would benefit from further investigation using a more established test that taps into a similar process, such as the EMT described earlier.

3.1.2.4 Ecological validity

Although content, response-process and convergent validity were specifically addressed in this study, it became apparent during the development of the test that ecological validity had not been determined. This refers to the extent to which the findings can be generalised to real life. On paper, the test may appear to be valid, but if a child scores highly on the FACES 2.0 test, does that mean they are able to decode facial expressions in real life? The teacher rating gave some support to the ecological validity but was limited in that it

involved judgment likely to be mediated by other factors such as personality and behaviour. Geher and colleagues identified the 20 highest achieving and 20 lowest achieving students on an EI measure and asked all 40 to watch a videotape of other students talking about what was on their minds. The high scoring EI group were significantly better able to identify the feelings of the students from the recordings than the low scoring group (Geher et al., 2001). This gives some evidence that a test situation can generalise to real life, although further investigation of the ecological validity of the FACES scale would add weight to the findings.

3.1.3 Further properties of the scale

The FACES 1.0 test was designed to identify children with deficits in facial expression decoding. To determine whether this had been achieved, it was necessary to examine facility values and the item-map. The findings from Trial One were encouraging and suggested that the FACES 1.0 test covered a range of pupil ability. The item map identified a group of pupils who had low scores on the test. These findings were replicated in the trial of FACES 2.0 and the teacher rating confirmed that the FACES 2.0 test was accurately identifying the lowest scoring group. It is possible that these pupils were not able to decode facial expressions but it is equally possible that they were not able to access the format of the test which was not picked up during observations. It is unlikely that a teacher would make a judgment about a pupil's ability based solely on the results of one test, but rather the teacher might choose to investigate further by informal interview with the child to find out whether the test was accessed correctly. If it was, the FACES 2.0 score may prompt a discussion which draws together observations from other sources and might include the pupil's relationships with peers and behaviour.

At the least, the identification of this group may help teachers to understand the difficulties experienced by the children which may reflect on their relationships and, as a result, may put them at risk of negative outcomes. At the extreme, the literature revealed a link between facial expression decoding and autistic spectrum disorders. The DSM-IV™ gives one of the diagnostic criteria for an autistic spectrum disorder as "*marked impairment in the use of multiple nonverbal behaviours such as eye-to-eye gaze, facial expression, body postures and gestures to regulate social interaction*"(American Psychiatric Association, 1994). FACES may contribute to the early identification of such disorders.

It is important to note here that the FACES 2.0 score alone would not be used in isolation. It would be intended to help identify pupils at risk as part of a pupil profile that covers a range of information on children including academic, personal and social factors, and home background.

FACES did not appear to discriminate well among the most able pupils. As the aim of the test was to identify those with deficits, this was not a cause for concern. Indeed, if, as

Baron-Cohen suggests, ability to decode facial expressions stabilises around the age of 5, this ceiling could reflect maturation of the skill.

Principal components analysis found a high level of unexplained variance. The Rasch model would assert that the unexplained variance can be accounted for by random noise and that, as this is evenly distributed throughout the test, is not degrading to test performance. It was not possible to evaluate this finding against those of the other instruments as variance was not reported in those studies.

The item-map suggested that, although the emotions appeared to be relatively well distributed across the difficulty range, 'happy' appeared to be an easier expression to decode. This is reflected in another reading of the literature. Markham and Adams (1992) found that four year olds were as able to identify the happy expression as seven year olds and these findings were supported by Widen and Russell (2008). In Markham and Adams' study, however, there was evidence that recognition of each expression was acquired incrementally in the following order; happy, sadness, anger, fear, surprise, disgust. There was no evidence from the FACES trial to support this.

3.2   Can the instrument be developed to minimise the confounding effects of language?

The findings from both the first and second trials suggested that the assessment was accessible for the age of the children and, once initial instructions were given, the children were able to understand what was required of them. Overall, children were able to interpret a vignette and select an appropriate answer from among the available responses without further reliance on language.

The instruments evaluated in the literature review all had a higher dependence on the use of language. The EMT and LEAS-C had a high language load both in the stimuli and, in the case of the LEAS-C, required some sophistication in emotional language to be able to provide the open-ended responses required. The reliability and validity of these measures, however, was shown by the literature to be good. The concern would be in the performance of the test for children with poor vocabulary or for English language learners. If a test is to be appropriate for identification of deficits in all young children, it is vital that it is able to discriminate without the confound of vocabulary. The DANVA was the least reliant on language with children simply being asked to provide a verbal label to emotions perceived and this task had a Cronbach α of 0.88. This instrument was only validated for use with children aged between 6 and 10 years which the literature suggests may be too late to begin interventions.

Overall, this research suggested that, compared to existing instruments, the FACES 2.0 test had real potential as a viable language-reduced alternative for measuring EI in young children.

3.3 Can the instrument be developed to be attractive and engaging for young children?

The final research question was concerned with whether the test could be developed to be attractive and engaging for young children.

Trial One found that, although the children enjoyed the computerised delivery and were happy with the expression-matching paradigm, there were some doubts about the suitability of the artwork. In discussions, none of the children professed active dislike of the images but evidence from post-test interviews suggested there was room for improvement. For the second version of the test, the stimuli were redrawn in colour and with more appealing facial expressions. Importantly, children were involved in the content of the stimuli. Observations during Trial Two suggested that the stimuli were improved. None of the children appeared uncomfortable or commented negatively on the vignettes or expressions.

It is a matter for concern that the studies which reported on the EMT, LEAS-C and DANVA did not attempt to investigate whether the stimuli appealed to the children and did not address whether they were comfortable with the testing process. With the variety of test delivery methods and accessibility of affordable options for artwork, researchers should be in a good position to balance a rigorous, scientific approach with consideration and sensitivity for the population they are working with.

The LEAS-C and EMT tests were also developed for use with children, but the DANVA was adapted from an adult version. The DANVA and EMT used photographs of adults and children and the LEAS-C relied on verbal stimuli. The cartoon vignettes in the FACES 2.0 test were developed specifically to appeal to children. An additional advantage of using cartoons was the appropriateness of its use with children with specific disorders. Being asked to look at photographs of adult or child faces may not be an appropriate method for identifying those with certain deficits, particularly those with autistic spectrum disorders. This is important because individuals with autistic spectrum disorders are likely to exhibit those very deficits that FACES is trying to identify. Children with autism or Asperger syndrome find it difficult to attend to faces and may, therefore, be at a disadvantage in a test using photographs. Cartoons may prove to be more accessible for this sub-group, although further trialling with a clinical sample would confirm this.

The stimuli for FACES employed cartoons, not only to appeal to children, but also to be appropriate for cultural sensitivities. The character was designed to show no stereotypical ethnic attributes in order to avoid being more accessible to particular groups of children. The universality approach, of course, suggests that emotional facial expressions can be identified across any culture, and that is fully acknowledged, but perceptions are important. The test would need to be acceptable to teachers, parents and children across a variety of

ethnic and cultural backgrounds. The EMT, LEAS-C and DANVA instruments did not report on cultural sensitivities which made it difficult to contrast them against the FACES scale. Young children could not be expected to attend to the battery of sub-tasks making up the other instruments evaluated in this study and a busy school day would simply be unable to accommodate their use. The LEAS-C consisted of four sections, with one task being delivered in two parts with a break in between. Although the time taken for the tasks was not reported, this suggests the test involved lengthy participation. Administration times for the DANVA were not reported although, as with the LEAS-C the test consisted of a number of subsections and it would not be possible to deliver all within a school day without tiring the children. The EMT was the least demanding, comprising four parts and taking around 15 minutes per child. The trial of FACES 1.0 took around 10 minutes and observations suggested that the children were beginning to lose concentration well before the end of the test. It is clear, however, that none of the measures discussed made claims to extend their use outside academic research. The FACES 2.0 test was found to have a good reliability when compared with other similar tasks and, crucially, was found to take only 5 to 6 minutes. This would suggest it is appropriate for maintaining the attention of the majority of children.

The random delivery meant that some vignettes appeared twice in succession and although this puzzled the children, they were able to answer the questions appropriately on repeated presentation. In fact, it is possible that having some repetition allowed children to confirm their interpretation of the item which may have contributed to the reliability of the overall test.

The instrument utilised computerised delivery which proved to be effective for assessment and attractive and engaging for the children. It reduced administration to 5 or 6 minutes per pupil, it standardised delivery and minimised the need for teacher intervention. Overall it was considered to be promising as a quick screening tool in a classroom setting.

3.4   Limitations

It would be useful to undertake some further exploration of the FACES test. The current research did not include analysis of the differences in performance of the test between boys and girls. This would be interesting from a theoretical point of view but also in terms of test performance. Differential item functioning analysis would provide evidence to suggest whether the test was biased towards one gender. Additional work might also include a comparison of correctness as assigned by the researcher and correctness as judged by a consensus scoring method. The percentage of variance explained by the FACES scale was low and further investigation of this would be advisable.

Although convergent validity was addressed in the second trial using the teacher rating scale, further evidence would be needed in order to be confident that the test was, indeed, measuring EI as conceptualised within the framework. Ideally, a sample of children would

be tested using the FACES scale and the EMT, the only other instrument designed for use with a similar age group.

At the outset of this research, it was intended that an adaptive test be developed. In order to develop this, it would have been necessary to have a large item bank trialled by a large number of individuals and restraints of time and resource did not allow for this. However, the groundwork has been laid for an adaptive test which may further reduce the administration time per pupil.

FACES was specifically designed to address issues of language and culture. A language-reduced paradigm was employed to ensure equal access to the test by children with differing levels of English language acquisition. The cartoon stimuli were intended to reduce the Anglo-centricity often found in test stimuli. An assumption was made that this had been achieved but was based on a small sample of 170 children. Level of language skill and cultural background were not controlled for. This research would benefit from further work with children of differing language levels and from a culturally diverse sample. As this research has addressed the link between facial expression decoding and autistic spectrum disorders, it would also be pertinent to obtain data from a clinical sample.

3.5   Implications

This study has implications for education practice. If, as is suggested here, it is possible to identify a subset of children with poor facial expression decoding skills, and this was substantiated by other evidence that suggested the child were at risk of negative outcomes, interventions could be put in place. Research suggests that poor facial expression decoding is linked to autistic spectrum disorders. Whilst it is unlikely that autism would be undiagnosed by the age of 4 or 5, Asperger syndrome can be more problematic to identify. If the FACES test were to provide an impetus for further investigation, this could be very valuable in early identification. Of course, deficits in facial expression decoding are not limited to a clinical population. Some children may exhibit no cognitive or learning difficulties but may simply have problems with relating to others. As the literature has shown, there is evidence that children who are better able to relate to teachers and peers are less at risk of negative outcomes.

Identifying deficits is valuable but particularly so if interventions are available. Silver and Oakes (2001) describe a randomised, controlled trial which found a computer program (Emotion Trainer) to contribute to gains made by 12 to 18 year olds in the recognition and prediction of emotional responses in others. A further study found impaired recognition of facial expression to be improved in a group of individuals with schizophrenia (Marsh et al., 2007). Although evidence has been found for interventions leading to an improvement in facial expression decoding skill, further work would be needed to identify whether those improvements were transferred to real life and were maintained over an extended period.

The findings of this thesis may contribute modestly to the field of research into facial expression decoding and emotional intelligence. This research has built on the foundations of important work in the field of EI measurement to develop a new instrument with considerable potential for research with non-clinical samples of young children, including those with low verbal language skill. The findings have suggested that researchers can be confident in their use of the measure with the FACES 2.0 test properties found to be comparable with other key instruments. A central benefit for researchers is that FACES 2.0 is considerably less time consuming than other measures and can be administered quickly to large groups. This makes the test appealing for use as a quick, light-touch measure of EI. Indeed, it has been argued that brevity of measurement is fundamental to research for which ecological validity is an important consideration (Lane et al., 2009). It could be posited, then, that the FACES 2.0 test may give a more valid reflection of EI skill than some of the more detailed instruments. A further benefit for researchers is that the simplicity of the administration process does not require any high level of training or expertise.

There is a key consideration about the use of the FACES 2.0 test which has been referred to at several instances through this study and which bears repetition. Although the test has been found to successfully identify a low ability subgroup in facial expression decoding, it is absolutely clear that this information should not be used in isolation to make judgments about pupils. Analysis of statistical information suggests that, overall, the test would correctly identify those with deficits but no single test can be 100% correct. The test may fail to identify a pupil with low ability. It may incorrectly identify a high ability pupil as having low ability. This highlights the importance that the FACES 2.0 scores contribute to a sophisticated profile which takes into account other abilities, personal and social factors, cultural capital, age, gender and, importantly, the reliability of any measures used.

4 Conclusions

This research has found that the new FACES scale has promise as an assessment to measure facial expression decoding in young children which could contribute to a profile of pupil strengths and weaknesses. The test successfully identified a low performing subgroup of children at risk of negative outcomes, giving practitioners additional information to put interventions in place at an early stage, where appropriate.

The FACES 2.0 test was shown to be more reliable than existing instruments claiming to measure the same concept and as reliable as measures of EI relying more heavily on the use of language. There were strong indications of good content, convergent and response-process validity. Importantly, during this study it became apparent that ecological validity was equally important and that further work would be needed to determine whether ability on the test transfers to ability in real life.

FACES employed an innovative paradigm which successfully reduced the use of language while maintaining good reliability. The computerised format meant that the assessment was quick to deliver and standardised administration in a way that would be difficult using other methods. The testing required little teacher intervention and no teacher judgment which may have allowed for bias in the responses. FACES may provide some validation for teachers in judgments already made. No specialist knowledge or expertise was required to deliver the assessment.

Children were involved in the production of the stimuli for FACES 2.0 and it could be argued that this contributed to the improved reliability of the test. Whether it contributed or not, it made the test more appealing and engaging to its target audience, the value of which should not be underestimated. As a result, the testing process was more appropriate for the children and not imposed upon them by adult assumptions and interpretations. Overall, the children were comfortable with the picture matching paradigm and were engaged by the artwork.

FACES was found to successfully identify a low achieving subgroup within a short period of time and, taken with the other findings, this suggested that it was appropriate for use as a screening instrument as part of normal classroom practice.

In summary, although further work to validate the findings has been identified, this study has found the FACES scale to be a valuable tool for the identification of facial expression decoding in young children.

5 Appendices

5.1 Ethics proposal

Durham University

School of Education

Research Ethics and Data Protection Monitoring Form

Research involving humans by all academic and related Staff and Students in the Department is subject to the standards set out in the Department Code of Practice on Research Ethics. The Sub-Committee will assess the research against the British Educational Research Association's *Revised Ethical Guidelines for Educational Research* (2004).

It is a requirement that prior to the commencement of all research that this form be completed and submitted to the Department's Research Ethics and Data Protection Sub-Committee. The Committee will be responsible for issuing certification that the research meets acceptable ethical standards and will, if necessary, require changes to the research methodology or reporting strategy.

A copy of the research proposal which details methods and reporting strategies must be attached and should be no longer than two typed A4 pages. In addition you should also attach any information and consent form (written in layperson's language) you plan to use. An example of a consent form is included at the end of the code of practice.

Please send the signed application form and proposal to the Secretary of the Ethics Advisory Committee (Sheena Smith, School of Education, tel. (0191) 334 8403, e-mail: Sheena.Smith@Durham.ac.uk). Returned applications must be either typed or word-processed and it would assist members if you could forward your form, once signed, to the Secretary as an e-mail attachment

Name: Katharine Bailey   Course: MRes

Contact e-mail address: kate.bailey@cem.dur.ac.uk
Supervisor: Dr Christine Merrell and Dr Richard Remedios

Title of research project: The development of an assessment to identify deficits in facial expression decoding in young children

Questionnaire

		YES	NO	
1.	Does your research involve living human subjects?	✓		IF NOT, GO TO DECLARATION AT END
2.	Does your research involve only the analysis of large, secondary and anonymised datasets?		✓	IF YES, GO TO DECLARATION AT END
3a	Will you give your informants a written summary of your research and its uses?		✓	If NO, please provide further details and go to 3b
3b	Will you give your informants a verbal summary of your research and its uses?	✓		If NO, please provide further details
3c	Will you ask your informants to sign a consent form?	✓		If NO, please provide further details
4.	Does your research involve covert surveillance (for example,		✓	If YES, please provide further details.

	participant observation)?			
5a	Will your information *automatically* be anonymised in your research?		✓	If NO, please provide further details and go to 5b
5b	IF NO Will you explicitly give *all* your informants the right to remain anonymous?			If NO, why not?
6.	Will monitoring devices be used openly and only with the permission of informants?	✓		If NO, why not?
7.	Will your informants be provided with a summary of your research findings?		✓	If NO, why not?
8.	Will your research be available to informants and the general public without authorities restrictions placed by sponsoring authorities?	✓		If NO, please provide further details
9.	Have you considered the implications of your research intervention on your informants?	✓		Please provide full details
10.	Are there any other ethical issues arising from your research?		✓	If YES, please provide further details.

Further details

The study will be conducted in local primary schools with young children aged between 4 and 6. Verbal consent will be obtained from Head teachers with written consent obtained from parents. Children will be advised that participation is voluntary and they may withdraw from the research at any time.

Head teachers will be given a verbal explanation of the aims of the research and how we intend to use the data gathered. Children will be told verbally that they are participating in some research to help design a new test. They will not be given written information as that would be inappropriate for their age. They will not be told specific detail about the research aims as this would jeopardise the research design.

Data collected will include pupil names in order to match pupils with pre-existing data on maths and reading attainment. Once matching has been performed, names will be removed from the data set and replaced with a pupil identifier code.

When the research is completed, the Head teachers will be given a report on the findings which they may choose to share with pupils if they feel this is appropriate.

The intention is produced a test which is appropriate for identifying deficits in emotion recognition. As such it needs to be attractive and fun to engage the pupils. The research conducted in school will be conducted sensitively, the stimulus materials is colourful and enjoyable. Although the content does include emotional material, all stimulus material is in cartoon format with no scenes likely to cause distress or discomfort for the children.

Declaration

I have read the Department's Code of Practice on Research Ethics and believe that my research complies fully with its precepts. I will not deviate from the methodology or reporting strategy without further permission from the Department's Research Ethics Committee.

Signed …………………………………………...   Date: …………………………

SUBMISSIONS WITHOUT A COPY OF THE RESEARCH PROPOSAL WILL NOT BE CONSIDERED.

Attached:

Copy of head teacher consent form
Copy of parental consent letter
Research proposal

5.2   Access letter to school

Dear

Thank you for kindly agreeing to allow us to come into school on Monday 20th March. We greatly appreciate your help.

I will be coming with Gideon Copestake, another researcher here at PIPS. We have both had police criminal record checks and have clearance to work with children.

We will bring four laptop computers with us and aim to trial the questions on 4 children at a time. We are interested in testing children from the Reception and Year 1 classes and our aim is to work with as many children as we practicably can during the day. As requested, I am enclosing enough letters to go to the parents of all the Reception and Year 1 children, although we do not anticipate having enough time to assess all of them. I wonder if you would be able to let us have a room to work in?

We hope to arrive at about 9.30 am and look forward to meeting you then.

Kind regards,

Yours sincerely

Katharine Bailey
PIPS Project

5.3 Access letter for parents

Dear Parent

As part of a research project the CEM Centre, based at Durham University, is developing a test to see how well children can interpret the emotions of others. The head teacher of your school has kindly agreed to let us try out some questions with your child's class.

The questions are fun for the children to complete and no pressure will be put on them whatsoever. Two researchers will be visiting the school to try out the questions. Both have had full police criminal record checks and have clearance to work with children. The data we collect is subject to the Data Protection Act and will not be revealed to any third party. All data will be destroyed after 6 months.

Please let the class teacher know if you do not wish your child to be involved.

Many thanks.

Yours sincerely

Katharine Bailey
PIPS Project

5.4 Item bank for FACES 1.0

Item no	Picture	Target	Distracter 1	Distracter 2	Distracter 3
1	sad1	sad	happy	angry	disgust
2	angry1	angry	sad	neutral	happy
3	disgust1	disgust	angry	neutral	happy
4	fear1	fear	disgust	happy	neutral
5	happy1	happy	sad	sad	angry
6	surprise1	surprise	neutral	angry	disgust
7	sad2	sad	happy	angry	fear
8	angry2	angry	neutral	fear	happy
9	disgust2	disgust	happy	neutral	sad
10	fear2	fear	neutral	happy	disgust
11	happy2	happy	disgust	angry	sad
12	surprise2	surprise	fear	sad	neutral
13	sad3	sad	angry	happy	disgust
14	angry3	angry	sad	neutral	happy
15	fear3	fear	disgust	happy	angry
16	happy3	happy	surprise	sad	neutral
17	sad4	sad	happy	neutral	angry
18	sad1	very sad	happy	very angry	very disgust
19	angry1	very angry	very sad	neutral	happy
20	disgust1	very disgust	very angry	neutral	happy
21	fear1	very fear	very disgust	happy	neutral
22	happy1	happy	very surprise	very sad	very angry
23	surprise1	very surprise	neutral	very angry	very disgust
24	sad2	very sad	happy	very angry	very fear
25	angry2	very angry	neutral	very fear	happy
26	disgust2	very disgust	happy	neutral	very sad
27	fear2	very fear	neutral	happy	very disgust
28	happy2	happy	very disgust	very angry	very sad
29	surprise2	very surprise	very fear	very sad	neutral
30	sad3	very sad	very angry	happy	very disgust
31	angry3	very angry	very sad	neutral	happy
32	fear3	very fear	very disgust	happy	very angry
33	happy3	happy	very surprise	very sad	neutral
34	sad4	very sad	happy	neutral	very angry
35	sad1	sad	happy	very angry	very disgust
36	angry1	angry	very sad	neutral	happy
37	disgust1	disgust	very angry	neutral	happy
38	fear1	fear	very disgust	happy	neutral
39	happy1	happy	very surprise	very sad	very angry
40	surprise1	surprise	neutral	very angry	very disgust
41	sad2	sad	happy	very angry	very fear
42	angry2	angry	neutral	very fear	happy
43	disgust2	disgust	happy	neutral	very sad
44	fear2	fear	neutral	happy	very disgust

45	happy2	happy	very disgust	very angry	very sad
46	surprise2	surprise	very fear	very sad	neutral
47	sad3	sad	very angry	happy	very disgust
48	angry3	angry	very sad	neutral	happy
49	fear3	fear	very disgust	happy	very angry
50	happy3	happy	very sad	very sad	neutral
51	sad4	sad	happy	neutral	very angry
52	sad1	very sad	happy	angry	disgust
53	angry1	very angry	sad	neutral	happy
54	disgust1	very disgust	angry	neutral	happy
55	fear1	very fear	disgust	happy	neutral
56	happy1	happy	surprise	sad	angry
57	surprise1	very surprise	neutral	angry	disgust
58	sad2	very sad	happy	angry	fear
59	angry2	very angry	neutral	fear	happy
60	disgust2	very disgust	happy	neutral	sad
61	fear2	very fear	neutral	happy	disgust
62	happy2	happy	disgust	angry	sad
63	surprise2	very surprise	fear	sad	neutral
64	sad3	very sad	angry	happy	disgust
65	angry3	very angry	sad	neutral	happy
66	fear3	very fear	disgust	happy	angry
67	happy3	happy	surprise	sad	neutral
68	sad4	very sad	happy	neutral	angry
69	disgust3	disgust	happy	neutral	sad
70	angry4	very angry	very fear	neutral	happy
71	angry4	angry	sad	neutral	happy
72	sad1	very sad	happy	angry	very disgust
73	sad4	very sad	happy	neutral	surprise
74	angry1	angry	very fear	neutral	happy
75	fear1	fear	very angry	very happy	neutral
76	sad2	sad	neutral	surprise	fear
77	angry4	angry	very sad	neutral	happy
78	surprise2	very surprise	fear	very sad	neutral

5.5 Trial One cartoon vignettes
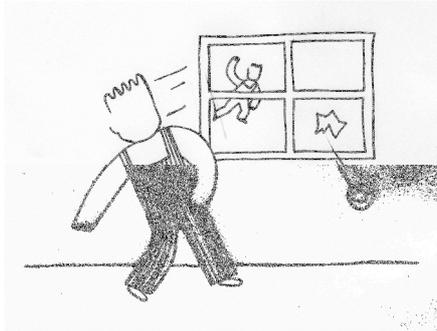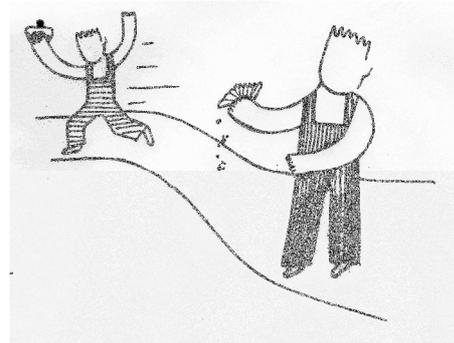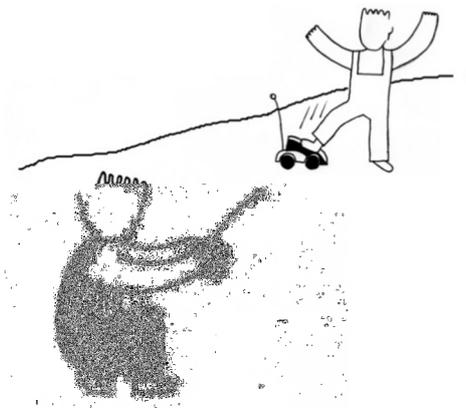

angry 1
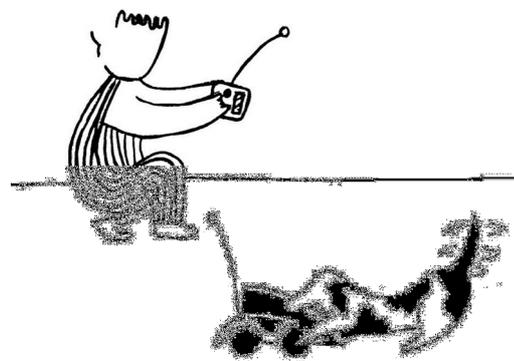
angry 2

angry 3

angry 4

disgust 1
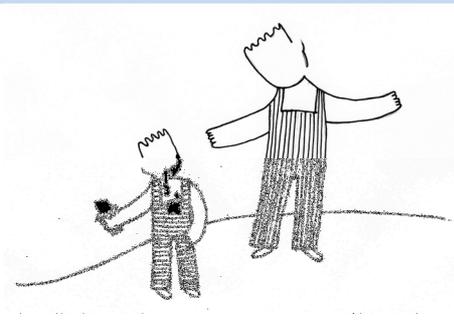
disgust 2
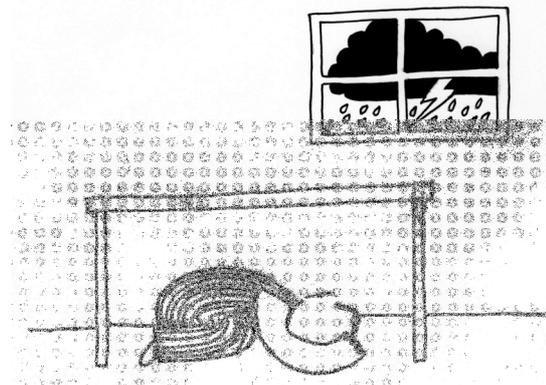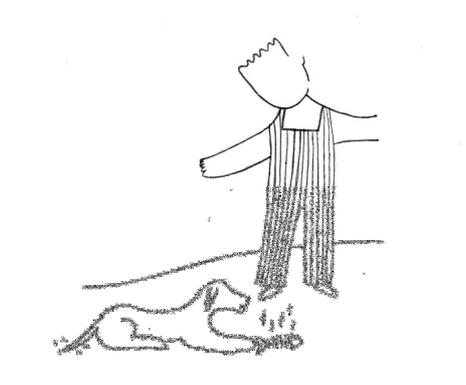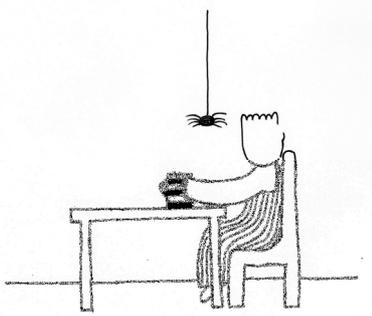
disgust 3	fear 1
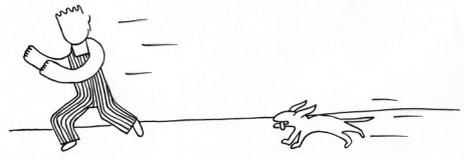
fear 2	fear 3
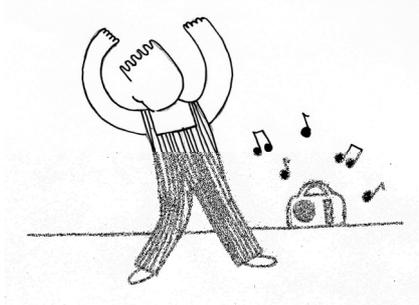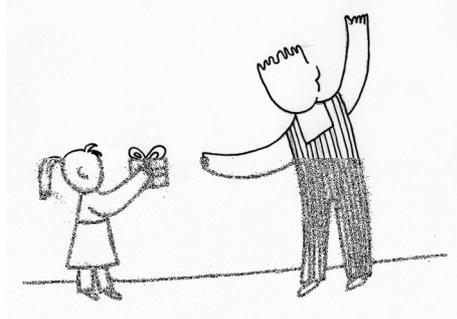
happy 1	happy 2
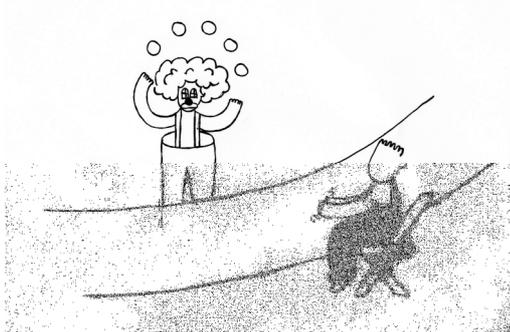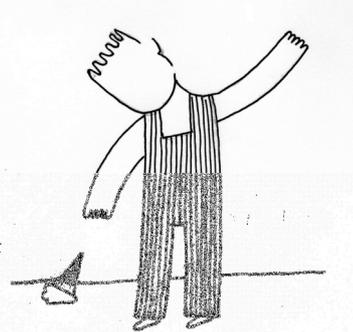
| happy 3	sad 1
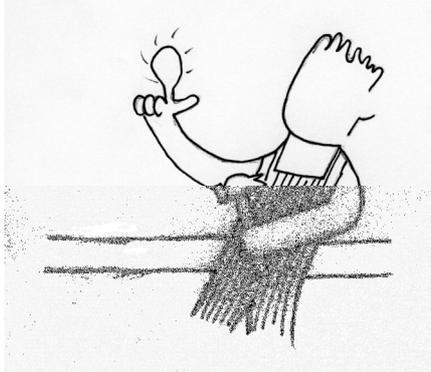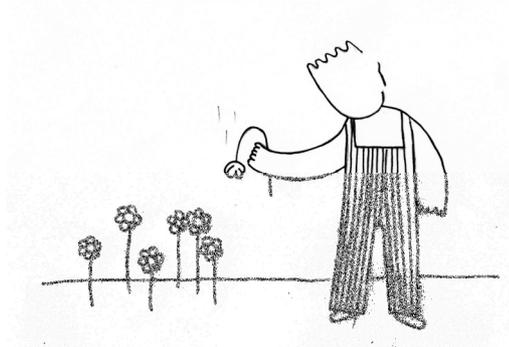
sad 2	sad 3
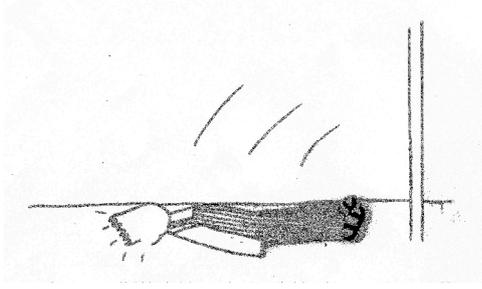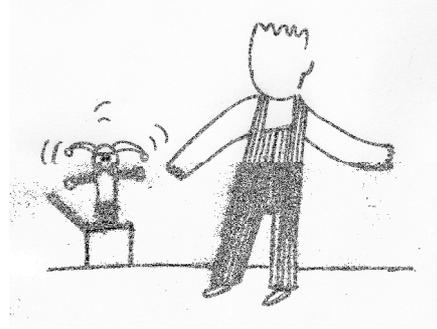
surprise 1
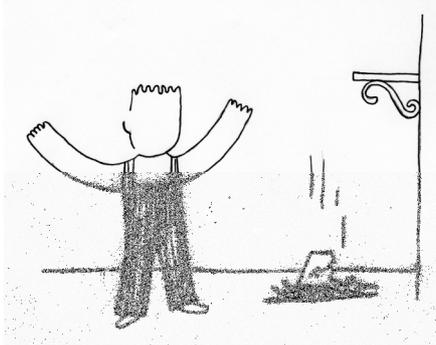
surprise 2

surprise 3

5.6   Trial One cartoon expressions
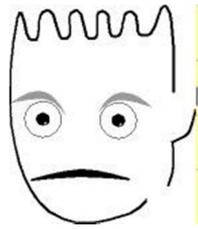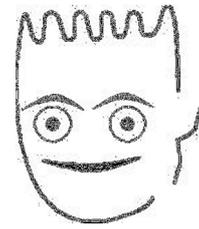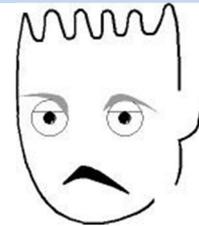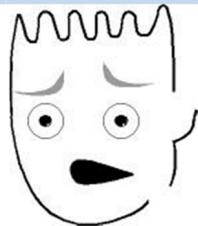
angry	disgust	fear	happy
sad	surprise	very angry	very disgust
very fear	very happy	very sad	very surprise
neutral			
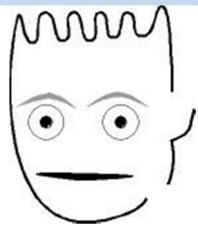
5.7 Item bank for FACES 2.0

| Item no	Picture	Target	Distractor1	Distractor2	Distractor3
1	angry1	angry	sad	neutral	happy
2	angry1	very angry	sad	neutral	happy
3	angry2	angry	neutral	fear	happy
4	angry2	very angry	neutral	fear	happy
5	angry3	angry	very fear	neutral	very happy
6	angry3	very angry	sad	neutral	happy
7	disgust1	very disgust	very angry	neutral	very happy
8	disgust1	disgust	very angry	neutral	very happy
9	disgust 2	disgust	happy	neutral	sad
10	disgust2	very disgust	happy	neutral	sad
11	disgust3	disgust	happy	neutral	sad
12	disgust3	disgust	very happy	very sad	very fear
13	fear1	very fear	very disgust	very happy	neutral
14	fear1	fear	very angry	very happy	neutral
15	fear 2	fear	neutral	happy	disgust
16	fear2	very fear	neutral	surprise	disgust
17	fear3	fear	disgust	happy	angry
18	fear3	very fear	very surprise	very happy	very angry
19	happy1	very happy	very surprise	very sad	very angry
20	happy1	very happy	surprise	sad	angry
21	happy2	very happy	very disgust	very angry	very sad
22	happy2	happy	very disgust	very angry	very sad
23	happy3	happy	surprise	sad	neutral
24	happy3	very happy	very surprise	very sad	neutral
25	sad1	sad	happy	angry	disgust
26	sad1	very sad	happy	angry	disgust
27	sad2	sad	happy	angry	fear
28	sad2	sad	neutral	surprise	fear
29	sad3	sad	angry	happy	disgust
30	sad3	very sad	angry	happy	disgust
31	surprise1	very surprise	neutral	very angry	very disgust
32	surprise1	surprise	neutral	very angry	very disgust
33	surprise2	very surprise	very fear	very sad	neutral
34	surprise2	surprise	very fear	very sad	neutral
35	surprise3	surprise	happy	neutral	sad
36	surprise3	very surprise	disgust	happy	fear

5.8 Trial Two cartoon vignettes
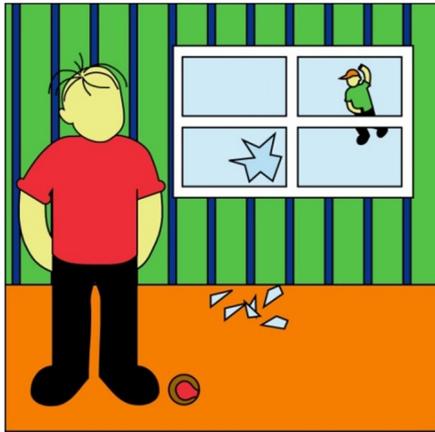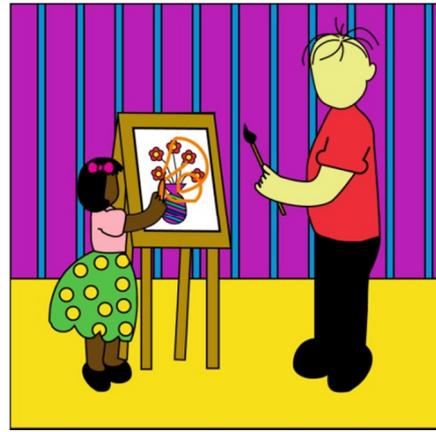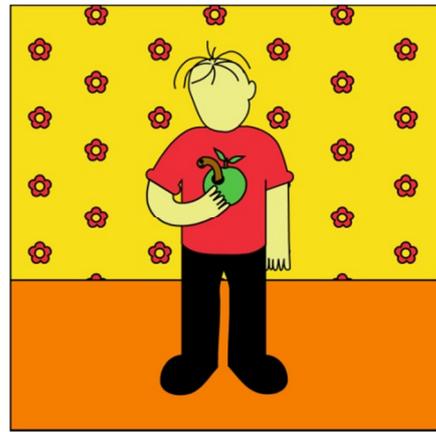

angry 1

angry 2

angry 3

disgust 1
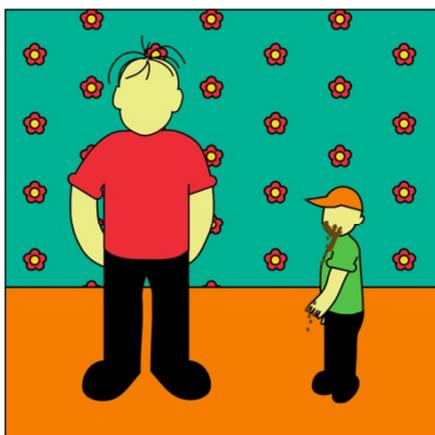
disgust 2
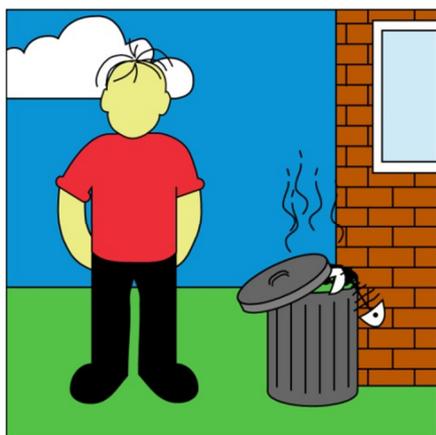

disgust 3

fear 1
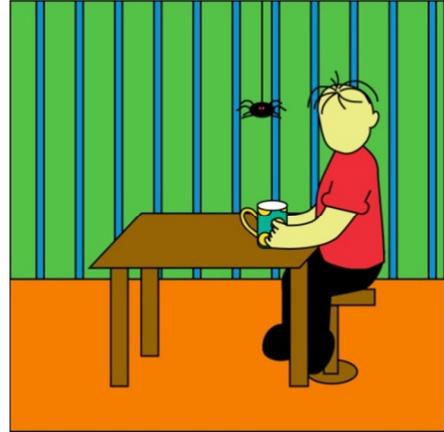
fear 2
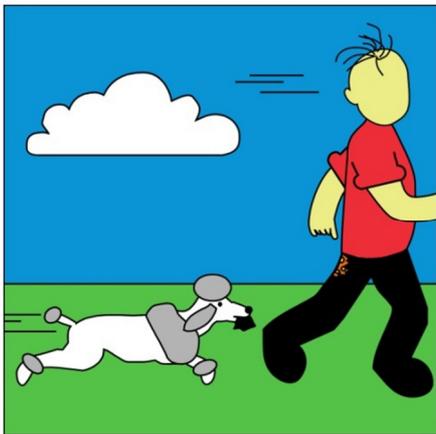
fear 3
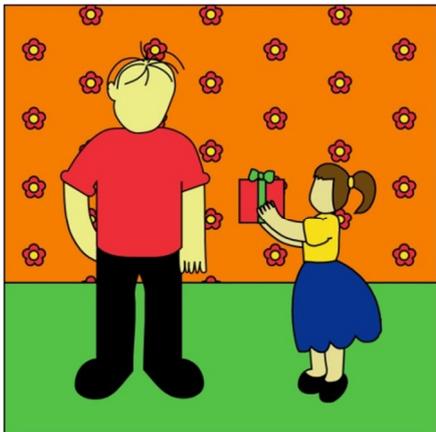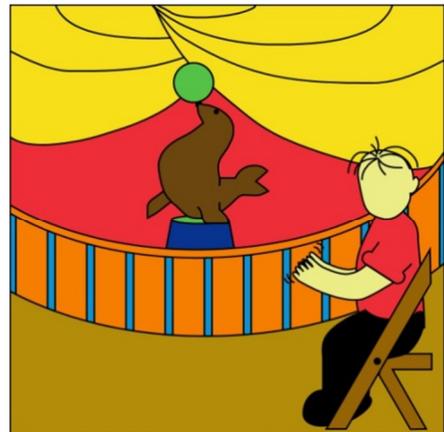
happy 1
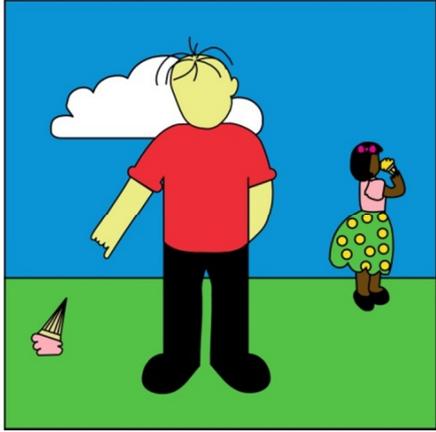
happy 2

happy 3

sad 1

sad 2
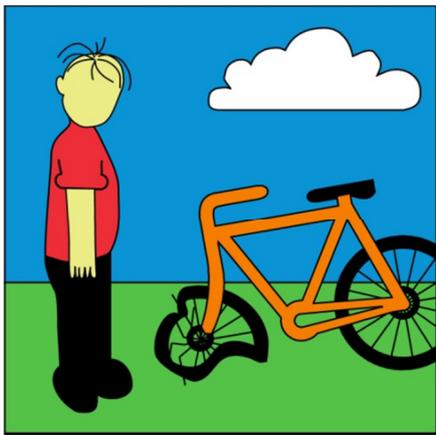
sad 3
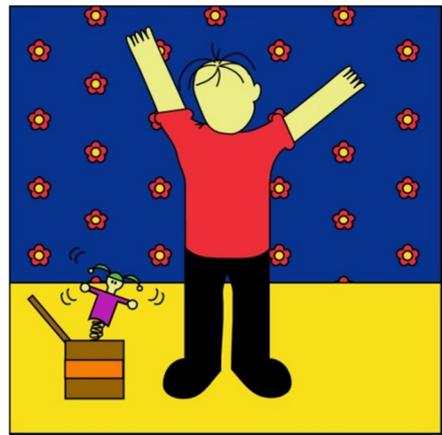
surprise 1

surprise 2

surprise 3

5.9 Trial Two cartoon expressions
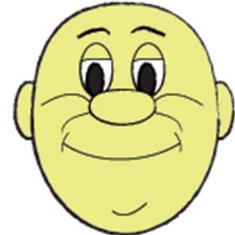
angry	disgust	fear	happy
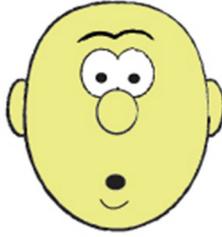
sad	surprise	very angry	very disgusted

very fear	very happy	very sad	very surprised
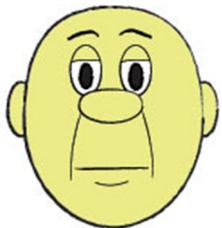
neutral			

5.10 Pupil biographical information and instruction screens

Please Enter The Pupils Details

Please Enter School Name — CEM

First Name — _____ Last Name — _____

Date of Birth — _____ Year Group — _____

Gender — _____ Today's Date — 01-Mar-2011

Next

Welcome to PIPS-Faces!

This computer program will show you some pictures of this man:

You will see him in each picture, but with his face missing. Look at each picture and think about how he might be feeling. Then look at the four faces next to the picture.

Choose which of those four faces you think belong to the man in the picture. You will see other people in some of the pictures but you need to think about how the MAN is feeling.

When you have decided, double click on that face. The computer program will take you to the next question.

When you ready, click the 'Go' button

Go

5.11 Data collection sheet for teacher rating

FACES Teacher Rating Scale

Class name _____

Class teacher _____

School _____

First name	Last name	Date of birth	Teacher rating

Teacher rating: 1 = one of the five pupils in the class least able to recognise emotions in others, 2 = not identified, 3 = one of the five pupils in the class most able to recognise emotions in others.

6 References

2007. Research Ethics Framework. Swindon: Economic & Social Research Council.

2011. *Ensuring Sound Conduct in Research* [Online]. Durham University. Available: http://www.dur.ac.uk/resources/hr/policies/research/ensuringsoundconduct.pdf [Accessed 12 February 2011].

AMERICAN PSYCHIATRIC ASSOCIATION 1994. *Diagnostic and Statistical Manual of Mental Disorders (DSM-IV™),* Washington, DC, American Psychiatric Association.

ARRINDELL, W. A. & VAN DER ENDE, J. 1985. An empirical test of the utility of the observations-to-variables ratio in factor and components analysis. *Applied Psychological Measurement,* 9.

ATTWOOD, T. 2000. Strategies for improving the social integration of children with Asperger syndrome. *Autism,* 4.

BAJGAR, J., CIARROCHI, J., LANE, R. & DEANE, F. P. 2005. Development of the Levels of Emotional Awareness Scale for Children (LEAS-C). *British Journal of Educational Psychology,* 23.

BARON-COHEN, S., RIVIERE, A., FUKUSHIMA, M., FRENCH, D., HADWIN, J., CROS, P., BRYANT, C. & SOTILLO, M. 1996. Reading the Mind in the Face: A Cross-cultural and Developmental Study. *Visual Cognition,* 3.

BLAIR, C. 2002. School Readiness: Integrating Cognition and Emotion in a Neurobiological Conceptualisation of Children's Functioning at School Entry. *American Psychologist,* 57**,** 111-127.

BOEHM, J. K. & LYUBOMIRSKY, S. 2008. Does happiness promote career success? *Journal of Career Assessment,* 16**,** 101-116.

BOND, T. G. & FOX, C. M. 2007. *Applying The Rasch Model,* New Jersey, Lawrence Erlbaum Associates.

BOWLBY, J. 1999. Attachment and loss. *Attachment.* 2nd ed. New York: Basic Books.

BRACKETT, M. A., RIVERS, S. E. & SHIFFMAN, S. 2006. Relating Emotional Abilities to Social Functioning: A Comparison of Self-Report and Performance Measures of Emotional Intelligence. *Journal of Personality and Social Psychology,* 91.

CHERNISS, C. 2004. Intelligence, Emotional. *Encyclopedia of Applied Psychology,* 2**,** 315-319.

CHERNOFF, H. 1973. The use of faces to represent points in k-dimensional space graphically. *Journal of the American Statistical Association,* 68.

CLEMENTS, D. H., SARAMA, J. H. & LIU, X. H. 2008. Development of a measure of early mathematics achievement using the Rasch model: the Research-Based Early Maths Assessment. *Educational Psychology,* 28**,** 457-482.

DARWIN, C. 1998. *The Expression of Emotions in Man and Animals,* London, Harper Collins.

DAWSON, M. E., SCHELL, A. M. & BOHMELT, A. H. (eds.) 1999. *Startle Modification,* Cambridge: Cambridge University Press.

DCSF. 2008. *The Early Years Foundation Stage Statutory Framework* [Online]. Department for Children, Schools and Families. [Accessed 23 July 2008].

DEPARTMENT FOR EDUCATION. *Achievement and attainment tables* [Online]. [Accessed 28 February 2011].

DEPARTMENT FOR EDUCATION. 2007. *Primary school (Key Stage 2) achievement and attainment Tables* [Online]. Available: www.education.gov.uk/egi-bin/performancetables [Accessed 28th February 2011].

EKMAN, P. 1973. Universal facial expressions in emotion. *Studia Psychologica,* 15.

EKMAN, P. & OSTER, H. 1979. Facial Expressions of Emotion. 528-554.

EKMAN, R. & FRIESEN, W. V. 1976. *Pictures of facial affect,* Paolo Alto, CA, Consulting Psychologists Press.

FARRONI, T., MENON, E., RIGATO, S. AND JOHNSON, M.H. 2007. The perception of facial expressions in newborns. *European Journal of Developmental Psychology,* 4**,** 2-13.

GEHER, G., WARNER, R. M. & BROWN, R. S. 2001. Predictive validity of the emotional accuracy research scale. *Intelligence,* 29.

GOLAN, O., BARON-COHEN, S., HILL, J.J. AND GOLAN, Y. 2006. The "Reading the Mind in Films" task: Complex emotion recognition in adults with and without autism spectrum conditions. *Social Neuroscience,* 1**,** 111-123.

GOLEMAN, D. 1995. *Emotional intelligence,* New York, NY England, Bantam Books, Inc.

HE, Q. 2006. CADATS. Durham: CEM, Durham University.

HEISE, D. R. 1985. Facial expression of emotion as a means of socialisation. *Electronic Social Psychology* [Online]. Available: www.indiana.edu/~socpsy/papers/FaceEmotionSocialization.html [Accessed 29 May 2011].

HERBA, C. & PHILLIPS, M. 2004. Development of facial expression recognition from childhood to adolescence: Behavioural and neurological perspectives. *Journal of Child Psychology and Psychiatry,* 45.

HERTENSTEIN, M. J. & CAMPOS, J. J. 2004. The retention effects of an adult's emotional displays on infant behavior. *Child Development,* 75.

HOBSON, P. R. 1986. The autistic child's appraisal of expressions of emotion. *Journal of Child Psychology and Psychiatry,* 27.

IZARD, C. E. 2001. Emotional intelligence or adaptive emotions? *Emotion,* 1**,** 249-257.

KEMPER, T. D. 1981. Social Constructionist and Positive Approaches to the Sociology of Emotions. *American Journal of Sociology,* 87.

KINGSBURY, G. G., MCCALL, M. & HAUSER, C. 2009. Tools for measuring academic growth. *Journal of Applied Measurement,* 10**,** 97-116.

KOHN, M. & ROSMAN, B. L. 1973. Cognitive functioning in five-year-old boys as related to social-emotional and background-demographic variables. *Developmental Psychology,* 8**,** 277-294.

KUCHUK, A., VIBBERT, M. & BORNSTEIN, M. H. 1986. The perception of smiling and its experiential correlates in three-month-old infants. *Child Development,* 57.

KWON, Y.-I. 2002. Changing Curriculum for Early Childhood Education in England. Access ERIC: FullText. *Early Childhood Research & Practice,* 4.

LANE, A. M., MEYER, B. B., DEVONPORT, T. J., DAVIES, K. A., THELWELL, R. C., GILL, G. S. & WESTON, N. 2009. Validity of the emotional intelligence scale for use in sport. *Journal of Sports Science and Medicine,* 8.

LINACRE, J. M. 2011. WINSTEPS® Rasch measurement computer program User's Guide. Beaverton, Oregon: Winsteps.com.

LUDEMANN, P. & NELSON, C. A. 1988. Categorical representation of facial expressions by 7-month-old infants. *Developmental Psychology,* 24.

MARKHAM, R. & ADAMS, K. 1992. The effect of type of task on children's identification of facial expressions. *Journal of Nonverbal Behavior,* 16.

MARSH, A. A., KOZAK, M. N. & AMBADY, N. 2007. Accurate Identification of Fear Facial Expressions Predicts Prosocial Behaviour. *Emotion,* 7.

MARSH, A. A., KOZAK, M.N. AND AMBADY, N. 2007. Accurate Identification of Fear Facial Expressions Predicts Prosocial Behavior. *Emotion,* 7**,** 239-251.

MATSUMOTO, D. & WILLINGHAM, B. 2009. Spontaneous Facial Expressions of Emotion of Congenitally and Noncongenitally Blind Individuals. *Journal of Personality and Social Psychology,* 96**,** 1-10.

MATSUMOTO, D., WILLINGHAM, B. & OLIDE, A. 2009. Sequential Dynamics of Culturally Moderated Facial Expressions of Emotion. *Psychological Science,* 20.

MAYER, J., ROBERTS, R. & BARSADE, S. 2008. Human abilities: Emotional intelligence. *Annual Review of Psychology,* 59**,** 507-536.

MAYER, J. D., SALOVEY, P. & CARUSO, D. R. 2004. Emotional Intelligence: Theory, Findings, and Implications. *Psychological Inquiry,* 15**,** 197-215.

MEDICUS, G., SCHLEIDT, M. AND EIBL EIBESFELDT, I 1994. Universal time constancy in movements of deaf-blind children. *Nervenarzt,* 65.

MERRELL, C. & TYMMS, P. 2007a. Identifying reading problems with computer-adaptive assessments. *Journal of Computer Assisted Learning,* 23**,** 27-35.

MERRELL, C. & TYMMS, P. 2007b. What children know and can do when they start school and how this varies between countries. *Journal of Early Childhood Research,* 5**,** 115-134.

MILES, S. B. & STIPEK, D. 2006. Contemporaneous and Longitudinal Associations Between Social Behavior and Literacy Achievement in a Sample of Low-Income Elementary School Children. *Child Development,* 77**,** 103-117.

MONTAGNE, B., VAN HONK, J., KESSELS, R. P. C., FRIGERIO, E., BURT, M. & VAN ZANDVOORT, M. J. E. 2005. Reduced efficiency in recognising fear in subjects scoring high on psychopathic personality characteristics. *Individual Differences,* 38.

MORGAN, J. K., IZARD, C. E. & KING, K. A. 2010. Construct Validity of the Emotion Matching Task: Preliminary Evidence for Convergent and Criterion Validity of a New Emotion Knowledge Measure for Young Children. *Social Development,* 19.

NELSON, C. A. 1987. The Recognition of Facial Expressions in the First Two Years of Life: Mechanisms of Development. *Child Development,* 58.

NOWICKI, S. & DUKE, M. P. 1994. Individual differences in the nonverbal communication of affect: The Diagnostic Analysis of Nonverbal Accuracy Scale. *Journal of Nonverbal Behavior,* 18**,** 9-35.

PARKER, J. D. A., CREQUE, R. E., BARNHART, D. L., IRONS HARRIS, J., MAJESKI, S. A., WOOD, L. M., BOND, B. J. & HOGAN, M. J. 2004. Academic achievement in high school: does emotional intelligence matter? *Personality and Individual Differences,* 37.

PARR, L. A., WALLER, B. M. & HEINTZ, M. 2008. Facial expression categorization by chimpanzees using standardized stimuli. *Emotion,* 8**,** 216-231.

PLUTCHIK, R. 1962. *The emotions: Facts, theories, and a new model,* New York, Random House.

POWER, M. J. & DALGLEISH, T. 1997. *Cognition and emotion; From order to disorder.,* Hove, UK, Pyschology Press.

QUALIFICATIONS AND CURRICULUM AUTHORITY 1999. Early Learning Goals. London: QCA.

ROSENTHAL, R., HALL, J. A., DIMATTEO, M. R., ROBERTS, P. L. & ARCHER, D. 1979. *Sensitivity to nonverbal communication: The PONS test.,* Baltimore, The Johns Hopkins University Press.

RUSSELL, J. A. 1995. Facial Expressions of Emotion: What Lies Beyond Minimal Universality? *Psychological Bulletin,* 118**,** 379-391.

SALOVEY, P. & MAYER, J. D. 1989. Emotional intelligence. *Imagination, Cognition and Personality,* 9**,** 185-211.

SCHOOL CURRICULUM ASSESSMENT AUTHORITY 1996. Nursery Education: Desirable outcomes for children's learning on entering compulsory education.   .

SCHUTTE, N. S., MALOUFF, J.M., HALL, L.E., HAGGERTY, D.J., COOPER, J.T., GOLDEN, C.J. AND DORNHEIM, L. 1998. Development and validation of a measure of emotional intelligence. *Personality and Individual Differences,* 25**,** 167-177.

SILVER, M. A. O., P. 2001. Evaluation of a new computer intervention to teach people with autism or Asperger syndrome to recognize and predict emotions in others. *Autism,* 5**,** 299-316.

TRENTACOSTA, C. J. & IZARD, C. E. 2007. Kindergarten Children's Emotion Compentence as a Predictor of Their Academic Competence in First Grade. *Emotion,* 7.

TYMMS, P. B. 2002. *Baseline Assessment and Monitoring in Primary Schools: Achievements, Attitudes and Value-added Indicators,* London, David Fulton Publishers Ltd.

VASILYEVA, M., LUDLOW, L. H., CASEY, B. M. & ST. ONGE, C. 2009. Examination of the psychometric properties of the measurement skills assessment. *Educational and Psychological Measurement,* 69**,** 106-130.

WATSON, D., CLARK, L. A. & TELLEGEN, A. 1988. Development and Validation of brief measures of positive and negative affect: The PANAS Scales. *Journal of Personality and Social Psychology,* 47.

WIDEN, S. C. & RUSSELL, J. A. 2008. Children acquire emotion categories gradually. *Cognitive Development,* 23.

YIEND, J. & MACKINTOSH, B. 2005. Cognition and Emotion. *In:* BRAISBY, N. & GELLATLY, A. (eds.) *Cognitive Pscyhology.* Milton Keynes: Open University.