The challenges of developing diagnostic assessments in countries rooted in rote learning – a case study around assessment in India

David Leach

A Research & Development Project

Submitted for the MSc Educational Assessment 2020

DEPOSIT AND CONSULTATION OF THESIS

One copy of your dissertation will be deposited in the Department of Education Library via the WebLearn site where it is intended to be available for consultation by all Library users. In order to facilitate this, the following form should be completed which will be inserted in the library copy your dissertation.

Note that some graphs/tables may be removed in order to comply with copyright restrictions.

Surname	Leach
First Name	David
Faculty Board	Education
Title of Dissertation	The Challenges of Developing Diagnostic Assessments in Countries Rooted in Rote Learning - A Case Study Around Assessment in India

Declaration by the candidate as author of the dissertation

1. I understand that I am the owner of this dissertation and that the copyright rests with me unless I specifically transfer it to another person.

2. I understand that the Department requires that I shall deposit a copy of my dissertation in the Department of Education Library via the WebLearn site where it shall be available for consultation, and that reproductions of it may be made for other Libraries so that it can be available to those who to consult it elsewhere. I understand that the Library, before allowing my dissertation to be consulted either in the original or in reproduced form, will require each person wishing to consult it to sign a declaration that he or she recognises that the copyright of this thesis belongs to me. I permit limited copying of my dissertation by individuals (no more than 5% or one chapter) for personal research use. No quotation from it and no information derived from it may be published without my prior written consent and I undertake to supply a current address to the Library so this consent can be sought.

3. I agree that my dissertation shall be available for consultation in accordance with paragraph 2 above.

Abstract

This thesis explores the challenges of developing diagnostic assessments to measure skills and competencies in an education system where assessments are rooted in a rote-learning culture. The research in this thesis applied the work of Marsh and Hau's 2007 research that issued a challenge to researchers that their methodology should be adopted in the analysis of substantive data. This challenge was applied to a 'real' (substantive) data set from a newly created diagnostic assessment in India, a country rooted in rote learning. The methodology describes a multi-stage process for exploratory factor analysis alongside practical cautions around limited use of rules-of-thumb, and considerations of missing data and interpretation of causality. The Marsh and Hau (2007) assertion is that the "polarization of substantive and methodological approaches to research and research training" must be reduced. The assessment at the centre of the research in this thesis was conceived, designed, developed and marketed by bringing together assessment implementors – experts, assessment developers, teachers, school leaders and examination board reviewers. This thesis concludes that the Marsh and Hau (2007) methodology is useful in a substantive research setting but argues that there is further polarisation to consider – that between researchers (methodological or substantive) and implementors of educational assessments.

Acknowledgements

My family, for supporting me throughout the last two years.

The Challenges of Developing Diagnostic Assessments in Countries Rooted in Rote Learning

A Case Study Around Assessment in India

Contents

Abstract 2
Acknowledgements 2
Contents 3
1. Introduction 4
   1.1. Background and Context 4
   1.2. Research Questions 5
2. Literature Review 7
   2.1. Introduction 7
   2.2. Moving Away from Pure Knowledge Testing 8
   2.3. Testing Culture 9
   2.4. From 'fast' Rote Learning to 'helpful' Diagnostic assessment 10
   2.5. Tools for Assessment Design Analysis 11
   2.6. Substantive Data 14
   2.7. Research Questions 15
3. Method 16
   3.1. Summary 16
   3.2. Context of the Study and Participants 16
   3.3. Instruments 16
   3.4. Data analysis 19
   3.5. School Related Effects 22
   3.6. Correlation of 'Scholastic' Performance with 'Psychometric' Evaluation 23
   3.7. Ethical considerations 23
   3.8. Summary 24
4. Results – Analysis and Presentation of the Data 25
   4.1. Data Analysis Procedures 25
   4.2. Demographic Data 25
   4.3. Descriptive Statistics 25
   4.4. Item Review – Classical Test Theory 28
   4.5. Item Review – Item Response Theory 30
   4.6. Factor Analysis 30
   4.7. Correlations between cognitive and psycho-educational constructs 39
   4.8. Summary 41
5. Discussion 44
6. Conclusions and Recommendations 53
   6.1. Conclusions 53
   6.2. Recommendations for Commercial Producers of Instruments and Exam Boards 56
   6.3. Recommendations for Further Research 57
   6.4. Research Limitations 58
7. Bibliography 60
8. Appendices 67
   Appendix A Paper Histograms 67
   Appendix B DISCA Competencies, Skills and Sub-skills 71

1. Introduction

1.1. Background and Context

The purpose of the research in this thesis was to gain insight into the challenges of developing assessments in countries where rote-learning is prevalent, and why, when there is so much published research and guidance, issues persist. The research explored the efficacy of assessment evaluation methodologies as applied to testing in an India context, by running quantitative analyses on a convenient secondary data set from an assessment designed to measure competencies and skills within subject contexts. The evaluation determined the construct validity of the assessment using published methodologies to formulate a series of recommendations for examination boards and agencies. The construct validity of a test is the validity of that test as a measure of real traits (Loevinger, 1957), and is described as having three aspects: a substantive component, a structural component and an external component. These three aspects relate closely to the three stages in test development: formation of an item pool, analysis and selection of the final pool, and correlation of scores with criteria and variables.

The work of Marsh and Hau was selected for key guiding principles due to their strong presence in educational and psychological research. Hebert W. Marsh has an h-index of 181 (Google Scholar, 2020b) and Kit-Tai Hau has an h-index of 59 (Google Scholar, 2020a). Their work on construct validity and related methods is highly cited, and one particular paper (Marsh & Hau, 2007) encourages the use of construct validation methodologies to be applied to substantive research. The paper from 2007 has been regularly cited, 220 times since publication (Google, 2020), and elaborates on construct validation methods. There are many methodologies describing construct validation methods (Flake, Pek, & Hehman, 2017), however the work of Marsh and Hau offers a practical Construct Validation Methodology and a challenge to researchers to explore the emergent methods against substantive data. They

recognise the significant developments in construct validation methods, but also challenge over-simplifications, or ‘rules of thumb’.

This thesis aims to bring together the work of Marsh and Hau and a conveniently available data set from an assessment sold commercially in India. The assessment under investigation was designed to measure competencies and high order thinking skills for children in India. In common with other countries, India has decided to participate in the international large-scale assessment of the Organisation for Economic Cooperation and Development (OECD), the Programme for International Student Assessment (PISA) (Banchariya, 2019; Singh, 2020). This will be the first time since 2009, when India ranked very low relative to other countries. The PISA assessment is designed to measure problem solving and higher order thinking skills (OECD, 2017), and participation can be a driver for change (OECD, 2019) within all aspects of education, including assessment systems. The educationalists in India have decided to try again with PISA indicating a readiness to measure their education system again and an interest in problem solving and other high order thinking skills. For a country rooted in rote-learning (Burdett, 2016) the challenge for the students will be new, and school leaders will want to know if their children have the skills to perform strongly on PISA. Against this backdrop, commercial organisations are more than willing to market assessments that target high order thinking skills (references withheld to maintain confidentiality) and claim to provide diagnostic reports.

1.2. Research Questions

This thesis explores the challenges of developing diagnostic tests, in a context more used to knowledge-based testing and rote-learning but where there is recognition and desire to encourage higher order thinking skills. The objective was to summarise a set of recommendations for examination boards and test developers, by addressing the following research questions:

1. *What degree of construct validity exists in tests designed for diagnostic purposes?*
    a. *How does a diagnostic assessment vary across schools, grades and subjects?*
    b. *What correlations in performance are apparent between cognitive and psycho-educational constructs?*
2. *How useful is the Marsh and Hau Construct Validation Methodology for a diagnostic assessment where rote-learning dominates the school culture?*

One commercial offering is DISCA – Diagnostic Interpretation of Skills and Competencies Assessment – which is described as an assessment solution designed to profile student personalities, their academic abilities, and to provide grading of competencies and skills. DISCA has extensive reporting systems and outputs for teachers and school leaders. DISCA was created by ABC Ltd, although both DISCA and ABC Ltd are pseudonyms for commercial anonymity and confidentiality reasons, and references to promotional materials, websites and identifying literatures have been withheld for similar reasons.

Data from a series of schools involved in a test session of DISCA in August 2019 was available and analysed to support exploring the research questions.

2. Literature Review

2.1. Introduction

Over the last 20 years there has been a growing trend in diverting assessments away from pure knowledge tests that rely heavily on rote-learning (memorisation) and towards assessments that test higher order thinking skills (Miri, David, & Uri, 2007; National Research Council, 1996). This shift has, in part, been driven by international benchmarking which has started to put more focus and value on these higher order skills (Burdett, 2016; Qadir et al., 2020). Countries are relying on immature assessment systems where new assessments are designed and created purely as an extension of the established approaches (G. T. L. Brown, 2011), rather than looking into new techniques which would likely be more beneficial to both students and teachers. The emergence of 'formative assessment' moves towards providing diagnostic data intended to help teachers (Black & Wiliam, 1998). However, these teachers are often ill-prepared to receive the data, and assessment providers are not used to gathering and reporting the data (Popham, 2011; Y. Xu & Brown, 2016).

Many countries, including India, Pakistan, Uganda and Nigeria, still have school cultures dominated by rote-learning (Browne, 2016; Burdett, 2017). For these systems and countries, studies exist that report on assessment quality based on analysis of the design of test items themselves as an editorial review, but few studies exist that explore empirical data from assessments undertaken by children. Whilst assessment design is well researched and understood, this literature review points to a lack of guidance for managing the change in the nature of assessments in countries where rote-learning dominates, as more diagnostic information is sought in favour over high-stakes grades and marks. Providing diagnostics information necessitates reporting against traits, or even simply a set of variables, that a teacher and student can relate to. This in turn requires that assessments are not only valid, but that the measured traits are known and described (Clarke, 2012; Miri et al., 2007).

Literature points towards a series of methodologies for analysing underlying traits and confirming the construct validity of an assessment (Marsh & Hau, 2007).

This literature review aims to discuss the challenges involved in designing valid assessments to give formative and diagnostic feedback for countries that are accustomed to using rote-learning throughout their education and assessment systems. It will be argued that simply reviewing test items from an editorial stance is not sufficient to provide validity evidence and instead construct validity methods (exploratory and confirmatory) are required. This view is supported in the work of Clark & Watson (1995) who map out a process for constructing validity and developing scales. Looking at the quality of assessment items by examining their content requires considerable experience and, even then, it is impossible to predict an item's efficacy by inspection. What is required is enough empirical and independent data of how items function in relation to other items, what latent traits emerge through correlation and consideration of how those latent traits could be described or named (Ferrer & McArdle, 2003).

2.2. Moving Away from Pure Knowledge Testing

There is growing interest in measuring students' personality, scholastic ability and 21st century competencies and skills in a graded manner, at many levels – across countries (Baird et al., 2011), within countries, within schools and for each student (Soland, Hamilton, & Stecher, 2013). It is possible that this ambition and interest harks back to the earliest days of measurement and the search for 'g', the general intelligence factor (Spearman, 1904), but continues to be apparent today. There has long been an ambition to create better integration between national and international measures, which are often summative, and more localised and formative (Wiliam, 2000). This ambition appears to be more towards a diagnostic approach and to inform formative measures rather than for certified or summative measurement (Soland et al., 2013). There appears to be many terms used to describe this

localised assessment – formative, classroom-based, continuous, -for-learning (Clarke, 2012) – but they all relate to the essence of the Black & Wiliam (1998) definition: “encompassing all those activities undertaken by teachers, and/or by students, which provide information to be used as feedback to modify the teaching and learning activities in which they are engaged”.

Although ‘diagnosis’ does not appear in this definition, it would seem that what is being described is a process of diagnosis, and therefore the process is diagnostic. The desire for all encompassing measures of children, in a terminal or summative form, over providing detailed insights to aid learning, varies in cycles over time and can be driven by political factors more than by fundamental educational objectives (Gove, 2014). As these desires come and go, so do the types of tests and the reporting approaches, and external accountability pressures distract the teaching and learning process. This distraction is potentially more extreme in the most challenged settings where accountability focus is most intense (Panesar-Aguilar, 2017). This leads other issues, such as recruitment challenges (Clotfelter, Ladd, Vigdor, & Diaz, 2004) and reviewing assessment data in unintended ways (Jennings, 2012).

2.3. Testing Culture

There are more and more tests and assessments appearing in emerging economies and these tend to be recall and rote learning oriented rather than for the testing of any higher order skills (Burdett, 2016, 2017). The work of Burdett investigates assessment design in India, Uganda, Nigeria, Pakistan and Alberta (Canada). One country that is seen as an “economic giant and potential global superpower” is India (Stambach & Hall, 2016). In India, the competition for places within schools systems that offer teaching by the most skilled teachers leads to a disproportionate focus on passing exams, and in turn concentration on rote learning. Stambach and Hall (2016) continue to describe the complex landscape of competition and settle on a word often used in their interviewing of students – fast. It is interesting to wonder whether ‘fast’ gets in the way of ‘thorough’, and, in educational terms,

is rote learning the *fast* way to progress? The authors suggest that attending to the compulsion and consequence of this situation will help us to understand better how to support the futures of children.

Commercial hunger drives a desire for companies to create assessments, and it can be too easy to mass produce poorly designed assessments to make revenues and profits. Exam boards and popular culture remain distracted by assessments that provide a grade rather than the diagnostic report (Browne, 2016). There is a need to evaluate what advice to give to exam boards on measuring the measurement – evaluating just how effective an assessment is.

2.4. From 'fast' Rote Learning to 'helpful' Diagnostic assessment

It might be argued that digitalisation of the classroom offers an escape from the rote learning trap and that artificial intelligence will provide the student performance analysis that teachers cannot (Agarwal, 2020). Agarwal makes the case that digital education could provide significant benefits to children in India but does not offer suggestions for changing the testing culture or supporting teachers with insights about their students. Students and teachers are not motivated to change, as any assessment that is not grade related is deprioritised in the minds of students and teachers alike (Warsi & Shah, 2019).

The benefits and possibilities of formative or diagnostic assessment has been widely researched (Black & Wiliam, 1998; Panesar-Aguilar, 2017; Stobart, 2008; Wiliam, 2000). Providing a useable and useful diagnostic assessment for teachers remains a target worth striving for, although we need to explore what can be provided to teachers and what are they seeking. Diagnostic information implies a view into a child that reveals something about their skills or competencies, but these terms have become muddled and entangled. Helpfully, the work of the National Center for Education Statistics provides a definition and hierarchy (Council of the National Postsecondary Education Cooperative, 2002). Their model lays out a foundation of traits and characteristics which develop through the learning process into skills,

abilities and knowledge. Further learning enhances these, and different combinations define competencies in individuals. This work is set within a context of competency-based education, and is more targeted at understanding post-secondary education, so it is less relevant to school-based assessment and feedback, but nonetheless serves as a good summary.

2.5. Tools for Assessment Design Analysis

Although diagnostic assessment is a valuable tool, it is not clear that teachers are ready to interpret the outputs of such assessments or that assessment providers produce valid diagnostic assessments. One working paper identifies that "assessment materials showed a very low proportion of higher-order skills" (Burdett, 2017, p. 5) with most rewards being received for rote-learning skills. The Burdett (2017) paper also suggests that basic assessment item quality is not present, and that assessments should be developed to allow students to demonstrate skills needed beyond schools, either for further education or employment. Studies into the India education system generally focus on economic and social parity factors (Deb, 2018), rather than the quality of assessments and measures used by schools and examination boards. Key to creating and reporting diagnostic information is designing assessments that report on specific and known constructs that a teacher can target with their teaching, and students can understand as they take ownership for their personal development (Shea & Duncan, 2013).

Within a context of schools in India, limited construct validation research exists and certainly little between traditional 'academic' subjects (maths, science, English) and psycho-educational constructs (adaptability, emotional intelligence). It appears that the academic subjects are sometimes referred to as 'scholastic' in literature pertaining to India. Areepattamannil (2014) investigated academic motivation and mathematics achievements as a comparison between India and Canada, although that study largely reported country-context differences.

There is a growing popularity in the use of latent variable modelling in psychological research in general and in educational psychology specifically (Liem & Martin, 2013). The psycho-educational constructs are in themselves unobservable and latent, so any measurement of these requires validation. Researchers have suggested (Marsh & Hau, 2007) that a *construct validation approach* could be adopted as a methodology for evaluating latent variable models. Although some will argue that construct validity has no basis in measurement theory and that it is simply a forced fitting of a theoretical concept of a data analysis (Colliver, Conlee, & Verhulst, 2012). This argument is specific to medical education and does not attempt to generalise further, but nonetheless represents an opposite position to that of Marsh and Hau.

Construct validity has many descriptions (AERA/NCME, 2014; Furr, 2018, p. 224) however the construct validation approach described by Marsh & Hau (2007) proposes that two significant modelling techniques exist within construct validation, *confirmatory factor analysis* (CFA) and *structural equation modelling* (SEM). The paper by Marsh and Hau (2007) provides a strong encouragement to researchers to adopt the construct validation approach, and outlines a number of approaches and interlocking principles. The paper does argue for the inclusion of multiple variables for each latent construct, and more than would eventually be used, with the assumption that the design of the final instrument will be based on factor structure analysis, which will in turn reduce the items in the measurement instrument.

The Marsh and Hau paper was written in 2007 and lamented the lack of "heuristic, non-technical demonstrations" of SEM, and it appears that in the last decade more and more articles aimed at the applied researcher have been published. More recent literature describes confirmatory factor analysis (CFA) and exploratory factor analysis (EFA) as two techniques that exist within the SEM hierarchy (Guo et al., 2019). The Guo et al. (2019) paper describes EFA as a technique for use where an instrument has not been fully analysed previously – as is the

case in this thesis. There is helpful guidance in the paper, although it does not meet the target set by Marsh and Hau for heuristic, non-technical demonstrations.

The techniques of CFA, EFA and SEM are now well documented in more practical guides and books (Byrne, 2012; R. B. Kline, 2015) and analysis methodologies are readily available within software applications (such as R and SPSS). One guiding article is that of Liem & Martin (2013) where it is summarised that factor analysis has three elements: (i) correlation between all latent factors (where there is more than one factor); (ii) each measure will have a (non-zero) loading onto the factor it aims to measure and a zero loading onto other factors; and (iii) uncorrelated error terms. It is also pointed out that SEM provides the structure of a predictive relationship between latent factors in a measurement. The use of CFA as applied to psychological research is questioned as over-simplistic (Marsh, Muthén, et al., 2009; Xiao, Liu, & Hau, 2019), and reinforces the benefit of exploratory approaches over confirmatory ones.

Marsh and Hau (2007) further elaborate positions on *missing data*, as well as *causality* and the *overuse of rule of thumb.* On missing data, they recommend against listwise and pairwise deletion and indicate a growing view that this is unacceptable. Instead, their recommendation is to consider the reason for missing data, randomness or variable related, and handle the imputation of data appropriately. The 'full-information maximum likelihood' algorithm is recommended for unbiased parameter estimates. On causality, Marsh and Hau (2007) remind that analysis can only show that data are consistent with predictions from causal models, rather than causality that has been proven. This was an important point to re-iterate through the research in this thesis. Finally, 'rule of thumb' approaches are criticised for their short-cut nature and over-simplification of interpretation, and the authors explore rules of thumb or 'golden rules' in other literature (Marsh, Hau, & Wen, 2009). This is helpful, but it does not fully recognise that a series of judgements need to be made in reviewing results, which necessarily require decisions about those results, which then affect subsequent analyses.

These ‘decisions’ seem to be little different to rules of thumb, except that the decisions are made within the context of the data, rather than transported in from generalisations in other research. This is a balance that needs monitoring carefully throughout.

2.6. Substantive Data

Let us remind ourselves of the comments of Stambach & Hall (2016) about India – an “economic giant and potential global superpower” – which raises the question: could education enhancement unlock that potential? If students knew what they had to develop to succeed, they could target their learning better and develop their skills more effectively. This thinking behind new assessments is being created by the regional office of ABC Ltd in India that produced a diagnostic assessment named DISCA. This assessment was used in schools in India in August 2019, when students at grades 5, 6, 7 and 8 were tested on English, mathematics, science and asked to complete a ‘Psychometric’ self-reporting questionnaire. The questionnaire was created against a framework that described a ‘personal and social’ construct consisting of: adaptability indices; emotional management; interpersonal; intrapersonal; and society. It is unusual to obtain measures of traditional ‘academic’ subjects, alongside self-reported personal and social survey information, particularly in assessment in India. This assessment mirrors some of what the international large-scale assessments aim to measure (Caro, Sandoval-Hernández, & Lüdtke, 2014; OECD, 2017).

The DISCA assessment is based on an assessment framework that elaborates the underlying latent traits that are intended to be measured, with each test item linked to the assessment framework and the latent traits. This provides a strong *a priori* definition of the factors or traits, and so was a relevant data source for the methodology of Marsh and Hau. The data represented significant number of students (N ≅ 7,000) each taking 3 papers plus completing a contextual questionnaire. The August 2019 data provide a rich source that Marsh and Hau

(2007) would call 'substantive' data, and it is convenient to use those data to examine their construct validation approach.

2.7. Research Questions

This thesis investigated test series from India and used the construct validation methodology to ascertain the robustness of the measures in the tests. Evidence for latent variables measuring higher-order skills was sought, and dependencies on personal and social aspects were investigated. Ultimately, this thesis aimed to show that illustrating constructs helps move focus away from the fast rote-learning habits and intense accountability measurement of teachers and schools' leaders, towards better diagnostics reporting and advice. The methodologies of the Marsh and Hau's (2007) Construct Validation Approach was adopted to critically test its applicability to a substantive data set. The research questions below were examined through analysis of secondary data – gathered in schools in India in August 2019 – which were grouped into four sets: English, mathematics, science and psychometric.

No	Research Questions	Methods	Analysis
1	***What degree of construct validity exists in a test designed for diagnostic purposes?***	A convenient data set from a commercial assessment design for students in India was analysed using research methods from Loevinger (1957), Marsh & Hau (2007) and Clark & Watson (1995)	Classical Test Theory and Item Response Theory was used to review and refine an item pool and enhance the reliability of test papers. Factor analysis was used to reveal dominant traits and correlations to intended construct design
1a	***How does a diagnostic assessment vary across schools, grades and subjects?***	The data were divided into school, age and grades and analysed for correlations and variability	Investigation of CTT, IRT and factor analysis data by ages and grades.
1b	***What correlations in performance are apparent between cognitive and psycho-educational constructs?***	The data included both cognitive measures and non-cognitive (psycho-educational) measures. The data was split to investigate correlations between the two	Linear regressions and graphical plotting of data
2	***How useful is the Construct Validation Methodology for a diagnostic assessment where rote-learning dominates the school culture?***	Review of full analysis outputs and interpretations to determine the overall efficacy of the methodology. The data set originates from work in a country where rote-learning is the system norm and a review of the results against known country characteristics	Retrospective summary of all data gathered, and analysis undertaken, in addressing research question 1

3. Method

3.1. Summary

The analysis of data in this project relied on the recommendations made by Marsh and Hau (2007) to use a Construct Validation Approach as a methodological approach in substantive studies. Marsh and Hau (2007) argue that the use of the Construct Validation Approach will bring together the methodological research with applied research to reveal methodological-substantive synergies. The analysis used data generated in an assessment in schools in India gathered in August 2019, so the study was largely a secondary data analysis. The purpose of this quantitative research was to explore the design efficacy of an assessment intended to be used as a student diagnostic, and rise to the challenge given by Marsh and Hau (2007) – for more substantive studies – and ultimately report whether their methodology yields insights that could enhance assessment validity, reliability and fairness.

3.2. Context of the Study and Participants

The data were collected through a series of instruments designed by ABC Ltd under a programme branded DISCA. The instruments were created against a series of constructs defined in assessment frameworks written by staff at ABC Ltd and a group of education experts within India. Data collection was undertaken in August 2019 in a group of 44 schools in India. DISCA is described as an assessment system intended to help students, teachers, parents and school administrators to profile students' personalities, academic ability, and grade competencies and skills. The assessment is designed for children in grades 5-8 (upper primary/middle schools at ages 10-14 years old).

3.3. Instruments

The instruments were a series of test papers across grades 5 to 8 in English, Mathematics, Science and Psychometrics – as these descriptions relate to named tests, it has been decided

that the capitalised form will be generally used. Psychometrics was the name given to a contextual questionnaire paper completed by all students of all ages. The three subject papers were referred to as the cognitive papers. The papers (instruments) were based on an assessment framework created specifically for the project developed in conjunction with a group of local education and assessment experts. The instruments went through a series of design, editing and review stages that initially created a bank of approximately 3,000 items. All the cognitive items were standard four-option multiple-choice questions, with a mixture of text, graphics, equations, graphs and other elements in the questions' stem.

The Psychometric paper was also a standard four option multiple-choice format, but the marking key allocated 0, 1, 2, 3 or 4 marks depending on the selection made. The editorial process for items involved staff experienced in assessment design, but the product owners did not readily accept assessment-oriented changes, instead basing their editing more on classroom content publishing standards. The test papers entered a Standards Setting phase in April 2017 and were used in three pilot sessions to prepare for commercial use. No data was available to ascertain how many pilots had been performed on each paper. The data used in this thesis was from the August 2019 session, and it can be seen, in Table 1, that the instrument had been through a good level of trialling before being used to formally report results.

Students				
Session Dates	**Standards Setting**	**Pilot Sessions**	**Paid Sessions**	**Total**
Apr-17	6,344			6,344
Jan-18		1,825		1,825
Aug-18		1,124	2,208	3,332
Jan-19		5,384	16	5,400
Aug-19			11,217	11,217
Dec-19			854	854
Total	**6,344**	**8,333**	**14,295**	**28,972**

Schools				
Session Dates	**Standards Setting**	**Pilot Sessions**	**Paid Sessions**	**Total**
Apr-17	13			13
Jan-18		11		11
Aug-18		12	10	22
Jan-19		20	1	21
Aug-19			47	47
Dec-19			11	11
Total	**13**	**43**	**69**	**125**

Table 1: Numbers of students and schools using the instruments to date

The tests were delivered in paper format under invigilation by school staff and marking of papers was undertaken by trained teams, coding against the marking keys. The assessments were designed to be undertaken as a set – English, Mathematics, Science and Psychometric – so that skills and competencies across the assessments could be reported. Diagnostic reports were provided to teachers and parents. There was a balance between analysis methodologies that worked within individual papers, and methodologies that analysed across groups of papers. The exact nature of this balance was explored in the analysis. The instruments were designed to measure competencies and skills articulated in assessment frameworks, and the competencies within the instruments were: communication, core thinking, creative thinking and personal & social. Each competency was described as a collection of skills and sub-skills, and competencies and skills were mapped across subjects and testing varied by both subject and age. The target competencies were mapped as shown in Figure 1. The competencies were further described by their supporting skills and sub-skills, with Figure 2 showing the skills (see Appendix B for the full list of sub-skills).

Subject	English				Mathematics				Science				Psychometric			
Grade	**5**	**6**	**7**	**8**	**5**	**6**	**7**	**8**	**5**	**6**	**7**	**8**	**5**	**6**	**7**	**8**
Communication	■	■	■	■	■	■	■	■	■	■	■	■				
Core Thinking	■	■	■	■	■	■	■	■	■	■	■	■				
Creative Thinking	■	■	■	■	■	■	■	■	■	■	■	■				
Critical Thinking	■	■	■	■	■	■	■	■	■	■	■	■				
Personal and social	■	■	■	■					■	■	■	■	■	■	■	■
Society	■	■	■	■					■	■	■	■	■	■	■	■

Figure 1: Competencies Measured by Subject and Age – grey signifies covered, white is not

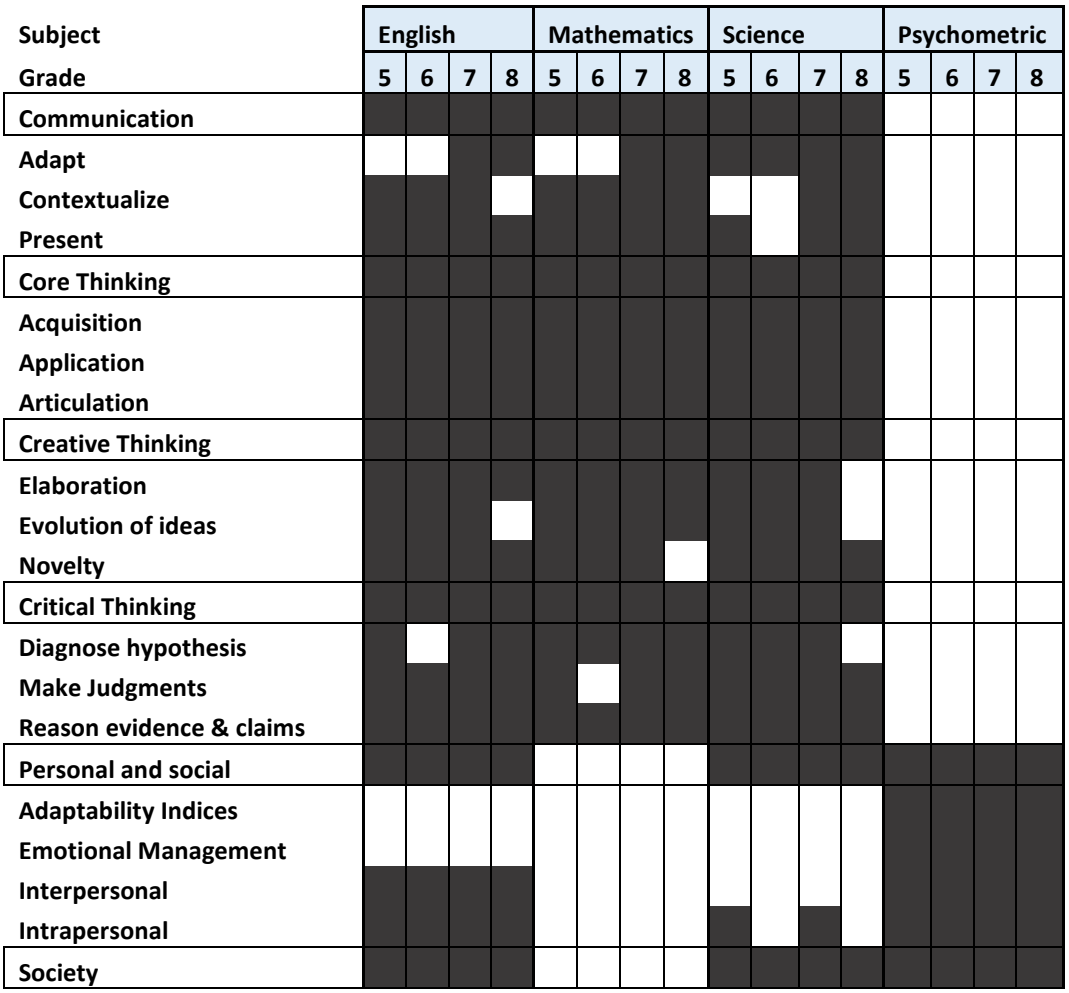


Figure 2: Competencies and skills Measured by Subject and Age – grey signifies covered, white is not

This breakdown of competencies and skills provided much detail in the assessment design assumptions to explore and confirm with the factor analysis methodologies of Marsh and Hau.

3.4. Data analysis

The data were in raw, anonymised form, which listed a school identifier, the paper taken, the student identifier, the question identifier and the response given. Other data were available in the data set, such as skill and competency descriptors for each test, and further class and enrolment identifiers for students. Importantly, the data provided were for the responses where a student had attempted the question, which meant that there was implied missing data. The word 'implied' is used here because there is an assumption that teachers and invigilators were using the instrument as intended – asking all students to attempt all questions, meaning that missing responses were treated as 'not attempted' or 'not reached'.

The data set was significant–just short of 700,000 observations–and this provided some processing speed limitations, but it was possible to sub-divide the data relatively easily. The majority of the analysis was carried out using packages with R (R Core Team, 2020) with support from generic office applications. Much time was spent understanding the data set, manipulating it into formats that could allow analysis, understanding peculiarities and idiosyncrasies, and finally readying data tables to allow simpler analysis. A whole series of R packages were investigated, tried, eliminated, used and learnt. The most useful non-standard R packages were: mirt (Chalmers, 2012), reshape (H Wickham, 2007), tidyverse (Hadley Wickham et al., 2019), openxlsx (Schauberger & Walker, 2020), psych (Revelle, 2018), cctICC function in the CTT R package (Willse, 2018) and nFactors (Raiche, Gilles and David Magis, 2020). The methodology used followed a multi-step process involving analysis, modification of items, instrument adjustment and iteration.

There were three test papers that contained dichotomous items (English, Mathematics and Science) and one paper that contained graded response items (Psychometric paper). The analysis for both types followed the same process but differed when item characteristic curves were being examined. The data analysis began with a straightforward review of the descriptive statistics to ensure that the population under test followed normal distributions. The data were reviewed for internal reliability and general central tendency reporting was undertaken.

As the test papers were designed as groups of questions to be reported as a total mark, classical test theory approaches were used to identify poorly performing test items, and these were targeted for removal. Inspection of the actual questions producing poor item means for test score groupings revealed confusing and badly worded questions, or answer choices. The worst items were identified, and their related data were removed from the data set to improve the internal reliability. Having reviewed the comparisons between commonly available IRT packages (Choi & Asilkalkan, 2019), the data were further examined through IRT

analysis, using the R mirt package (Chalmers, 2012). The fit parameters for M2, root mean square error of approximation (RMSEA), comparative fit index (CFI) and Standardized Root Mean Square Residual (SRMR) (Cai & Hansen, 2013; Dyer, Hanges, & Hall, 2005; R. B. Kline, 2015; J. Xu, Paek, & Xia, 2017) were examined to determine the best fitting models for the IRT analysis. The mirt package generated item characteristic curves (ICC) with expected and observed groupings. Inspection of these ICCs indicated more items that were performing poorly, and this provided a further group for removal. Test reliability was enhanced significantly by eliminating groups of items. The papers were evaluated through the KMO function in the R psych package and interpreted in relation to the Kaiser-Meyer-Olkin (KMO) Test (Kaiser, 1974), which confirmed the data were suitable for factor analysis (Kaiser, 1960), so the process was started. As the assessment had never benefitted from factor analysis previously, it was decided that exploratory techniques were preferable (Child, 1990).

Although the assessments were built on a hierarchy of competencies and skills (or latent traits), as in Appendix B, it seemed worthwhile first undertaking some exploration of the latent traits emerging from the performance of the test, items and response patterns, from a purely numerical and statistical stance. The literature around factor analysis is split on the approach of using Exploratory Factor Analysis (EFA) over Confirmatory Factor Analysis (CFA) (Orcan, 2018), but the essence seems to be that EFA is used where patterns are being explored and little or no factor development has occurred for the assessment generating the data. CFA is more useful for the confirmation of hypotheses and the validation of relatively well-developed variables (Child, 1990). There are several methods for proceeding with an exploration of factors (or Exploratory Factor Analysis) and for deciding on aspects of extraction, rotation and factor numbers (Costello & Osborne, 2005).

Where the data is relatively normally distributed, maximum likelihood analysis (ML) of factors is regarded as a good approach (Tucker & Lewis, 1973) and ML is used within the standard R

function, factanal. It was discovered early on that the data were normally distributed. A common method to begin the process or factor analysis and variable reduction, is to inspect a graphical representation that plots eigenvalues in descending order against factors, using a Cattell's scree plot (Cattell, 1966). A cut-off at an eigenvalue of one can then be applied to determine how many factors should be retained (Kaiser, 1960). Other methods are available, including parallel analysis and optimal coordinates (Raîche, Walls, Magis, Riopel, & Blais, 2013), which are proposed as more robust evaluations of the retained factor quantity. Research studies (Raîche, Riopel, & Blais, 2006; Ruscio & Roche, 2012) point towards the 'optimal coordinates' method as an effective and straightforward approach to implement for the evaluation of the number of factors.

The methodology for the factor analysis for this project was iterative. Firstly, the optimal coordinates number was determined for each paper, then the items loading most strongly on the number of factors was investigated. Secondly, the factor quantity was increased to reach a point where the hypothesis of perfect fit can longer be rejected. Thirdly, items loading least against the factors were eliminated and the investigation iterated until a pool of most representative items was identified. The variance explained by the reduced item pool was reported and evaluated. Investigations by individual paper (English, Mathematics and Science) were carried out initially and this was expanded to group papers by year groups of children.

3.5. School Related Effects

Research question 1a, related to the variability by groups, the investigation of school-related effects and understanding the relationship between schools and performance of students. A straightforward totalling of scores in English, Mathematics and Science papers by grade was used. This however required consideration as the CTT and IRT analyses indicated that a reduced item pool was preferential. There were no school-related characteristics in the data so only limited correlations were possible, other than simply school-to-school comparisons.

Some investigation of grade and subject correlations was possible and used to examine potential longitudinal mismatches in the measures.

3.6. Correlation of 'Scholastic' Performance with 'Psychometric' Evaluation

As explained, the assessment includes scholastic (cognitive) tests (English, Mathematics and Science) and a Psychometric test, which is akin to a personality survey and measures a series of psycho-educational constructs. The psychometric part of the test is intended to report against the competencies and skills of the assessment's framework (see Appendix B for details of the skills and competencies). In addressing research question 1b, the research in this thesis explored correlations in performance between the cognitive and psycho-educational constructs. The correlation between cognitive performance and competencies was explored through linear regression methods and graphical plotting (Fox, 1987; R Core Team, 2020). The assessment was designed to measure and report the competencies on separate scales, and regression models for these separate scales together with a single psychometric measure were explored.

3.7. Ethical considerations

The data used in this thesis was from sessions paid for by schools and related to student performance. As the data was from a commercial offering, there was some sensitivity about access to the data, and anonymity had to be maintained. The commercial owner of the instruments is not identified as this thesis could impact long term agreements with schools. The brand name of the instrument has also been hidden and all student and school identifiers were reduced to data tags that could not be tracked. The research work and data storage methods were approved by the University of Oxford Departmental Research Ethics Committee – reference ED-CIA-20-221. The output of the research was used to create this thesis, although the work may need to drive a high-level recommendation for the commercial entity that owns the instruments. This recommendation will not breach the Ethics Committee regulations.

3.8. Summary

The research in this thesis centred on evaluating construct validity and using methodologies suggested in publications. Many researchers cite Herbert W. Marsh in their work, and it is clear that he is a prolific researcher in this area, being cited very significantly, earning him a high h-index. Some of his work has been undertaken alongside Kit-Tai Hau, who also features as an often-cited researcher with a strong h-index. Amongst their many papers is one from 2007, cited over 200 times, which has been used as a central guiding methodology. In summary, this paper calls for *multiple* methods of analysis that should be evaluated against each other, but without forgetting the most basic of statistical principles and avoidance of generalisations. This guidance is elaborated further by Liem & Martin (2013) who propose that latent variable modelling techniques "have the capacity to answer this call" (p. 187). The Marsh and Hau methodology encourages the use of confirmatory factor analysis (CFA) and structural equation modelling (SEM). More recent literature describes CFA and EFA as two techniques that exist within the SEM hierarchy (Guo et al., 2019).

The final structure for the data analysis was guided overall by the Marsh and Hau (2007) methodology, with sub-processes designed around Loevinger (1957) and Clark and Watson (1995), to create the following process:

1. Evaluate test papers for reliability and identify any poorly performing test items. Remove items that perform poorly until a good reliability rating has been achieved.
2. Identify numbers of factors and factor grouping evident in the data.
3. Filter down to the items with the strongest factor loadings to create minimalist instruments.
4. Run factor analyses using both *a priori* competency and skills grouping and the newly revealed latent traits in 3 above.
5. Perform regression analyses especially in relation to school and age groupings.

4. Results – Analysis and Presentation of the Data

4.1. Data Analysis Procedures

The data analysed came from an August 2019 series used in schools in India as already described. These data were derived from test papers that were grouped into English, mathematics and science contexts, and were specifically designed for providing formative (diagnostics) feedback for children. The analysis approach used the methodology described by Marsh and Hau (2007) and submitted to their challenge of applying methodological research practices to substantive data research in order to seek out synergies. The test papers had been designed against an assessment framework developed by expert groups in India and were intended to measure various traits that could then be reported against for diagnostic purposes. No other confirmation of the efficacy of these test papers as assessment instruments had ever been carried out, so the analysis below had no *a priori* insights beyond the contents of the assessment framework. For that reason, a considerable refining of the assessment instrument was required before the research questions could be more specifically addressed.

4.2. Demographic Data

The data were available with some demographic variables included: school attended (anonymised), school year/grade and date of test. The students attended 44 different schools and grades 5, 6, 7 and 8 were represented. No gender data for students were available. All schools are in India and are a mixture of state and private schools and the data provided lacked any division to allow contexts to be investigated.

4.3. Descriptive Statistics

Overall inspection of results across papers showed that there were normal distributions in the data (see Appendix A). The results from each test paper were reviewed through generic

descriptive statistics reports, as shown in Figures 3 and 4. In general, these reports indicate a positive set of measurement instruments (test papers) that performed well. All test papers contained 35 multiple choice questions, each scored 1 or 0.

Test papers exhibited varying levels of difficulty, as evidenced by the range of means (11.00 to 21.21) versus a maximum score of 35. The proportion of missing data in the responses was 3.2% or below, and well below the 5% suggested as the maximum before more complex handling of the missing data (Graham & Hofer, 2000). From a visual scan of the responses, the missing values appear to be Missing At Random (MAR) (Chen, Wang, & Chen, 2012; Little, R. J. A., & Rubin, 2002). The missing values were replaced by 0 (incorrect) in all cases. The internal reliability across papers was shown to be reasonable but not strong in all cases.

Test paper P-5-SCN-A1-Y18 ($n = 1008$) showed students performing at an average of 11.0 ($s = 4.49$) with modest positive skewness and a leptokurtic distribution. The analysis for this paper showed a less than acceptable Cronbach's internal reliability ($\alpha = 0.66$).

Similarly, paper P-6-MATH-A1-Y-18 ($n = 1119$) student mean was 11.56 ($s = 4.65$) with modest positive skewness and a leptokurtic distribution. The reliability for this paper was less than acceptable ($\alpha = 0.68$).

All other test papers demonstrated low skewness and kurtosis, and acceptable or good internal reliability. The Standard Error of Measurement across all papers was relatively similar (2.54 to 2.70) indicating that student scores were within a ±5.4 scale score points at a 96% confidence interval.

	P-5-ENG-A1-Y-18	P-5-ENG-E-Y-18	P-5-MATH-A1-Y-18	P-5-MATH-E-Y-18	P-5-SCN-A1-Y18	P-5-SCN-E-Y-18	P-6-ENG-A1-Y-18	P-6-ENG-E-Y-18	P-6-MATH-A1-Y-18	P-6-MATH-E-Y-18	P-6-SCN-A1-Y-18	P-6-SCN-E-Y-18
Statistic	**P5 ENG**		**P5 MATH**		**P5 SCN**		**P6 ENG**		**P6 MATH**		**P6 SCN**	
items	35	35	35	35	35	35	35	35	35	35	35	35
N	1016	597	1015	596	1008	593	1120	1024	1119	1023	1118	1020
Missing values	2.0%	1.3%	2.6%	1.4%	3.1%	3.2%	2.4%	1.4%	2.9%	2.1%	2.5%	2.4%
Mean	15.40	21.21	12.90	17.16	11.00	14.28	14.32	19.55	11.56	14.36	12.51	15.63
Std.Dev	6.32	6.55	5.29	6.22	4.49	5.73	5.37	5.86	4.65	5.81	5.39	6.95
Min	3	2	2	1	1	1	4	4	0	2	1	0
Q1	11	16	9	12	8	10	10	15	8	10	9	10
Median	14	22	12	17	10	14	14	20	11	13	11	14
Q3	20	26	16	22	13	19	17	24	14	18	15	21
Max	33	35	30	33	30	29	32	34	31	33	32	35
Skewness	**0.59**	-0.21	**0.83**	0.15	**1.05**	0.29	**0.66**	-0.10	**0.98**	**0.65**	**0.87**	0.50
SE.Skewness	0.08	0.10	0.08	0.10	0.08	0.10	0.07	0.08	0.07	0.08	0.07	0.08
Kurtosis	-0.39	**-0.74**	0.43	**-0.61**	**1.37**	**-0.65**	0.11	**-0.58**	**1.37**	-0.03	**0.62**	**-0.62**
Alpha	0.82	0.85	0.76	0.83	0.66	0.78	0.75	0.81	0.68	0.80	0.75	0.85
SEM	2.68	2.54	2.61	2.55	2.63	2.66	2.70	2.59	2.62	2.63	2.68	2.66

Figure 3: Grade 5 and 6 Papers – Descriptive Statistics

	P-7-ENG-A1-Y-18	P-7-ENG-E-Y-18	P-7-MATH-A1-Y-18	P-7-MATH-E-Y-18	P-7-SCN-A1-Y-18	P-7-SCN-E-Y-18	P-8-ENG-A1-Y-18	P-8-ENG-E-Y-18	P-8-MATH-A1-Y-18	P-8-MATH-E-Y-18	P-8-SCN-A1-Y-18	P-8-SCN-E-Y-18
Statistic	**P7 ENG**		**P7 MATH**		**P7 SCN**		**P8 ENG**		**P8 MATH**		**P8 SCN**	
items	35	35	35	35	35	35	35	35	35	35	35	35
N	1193	416	1193	415	1192	415	1095	346	1095	346	1093	346
Missing values	2.1%	2.1%	2.1%	1.9%	2.6%	3.2%	1.4%	1.2%	1.8%	2.3%	2.0%	2.2%
Mean	13.63	20.00	14.66	19.59	14.57	20.58	15.11	20.80	15.07	20.09	13.02	18.92
Std.Dev	4.79	6.21	5.41	6.25	5.92	7.53	5.63	6.31	5.68	6.79	5.11	6.98
Min	1	2	2	0	0	1	1	3	3	2	0	2
Q1	10	16	11	16	10	15	11	16	11	15	9	13
Median	13	21	14	20	14	21	14	21	14	21	12	19
Q3	16	24	18	25	18	27	19	26	19	26	16	25
Max	33	33	32	33	34	34	34	33	32	33	30	33
Skewness	**0.67**	-0.41	**0.61**	-0.47	**0.56**	-0.24	0.48	-0.27	0.43	-0.31	**0.67**	-0.24
SE.Skewness	0.07	0.12	0.07	0.12	0.07	0.12	0.07	0.13	0.07	0.13	0.07	0.13
Kurtosis	**0.91**	-0.08	0.05	0.00	-0.19	**-0.85**	-0.09	**-0.69**	-0.40	**-0.67**	0.20	**-0.94**
Alpha	0.70	0.83	0.75	0.83	0.80	0.88	0.77	0.84	0.78	0.86	0.72	0.86
SEM	2.64	2.54	2.68	2.61	2.65	2.56	2.68	2.53	2.68	2.50	2.69	2.61

Figure 4: Grade 7 and 8 Papers – Descriptive Statistics

The data included results for the Psychometric test paper with ratings (0 to 4) on a series of statements aligned to psycho-educational constructs. The design of the paper implied that a

higher mark on the rating scales indicated a positive characteristic for the student, so overall totalling and descriptive statistical analysis has some merit but is treated with caution. The data for the psychometric tests included missing responses at a rate of <5% is all cases bar one. The test item PSY-40 had 7% of the data missing. The student (n = 6680) mean was 60.8 (s = 9.55) with negative skewness, -1.59, and a leptokurtic kurtosis of 6.32.

4.4. Item Review – Classical Test Theory

All test items were evaluated using the cctICC function in the CTT R package (Willse, 2018) and each item characteristic curve was reviewed using the guidance of T. Kline, 2005, p. 98. The item characteristic curves examined were plots of item means against percentile groupings. Many item characteristic curves indicated items that performed well, by discriminating in line with student overall test scores. However, several items performed poorly. Items i10115, i10733 and i30068 are examples of poor items, see Figure 5.

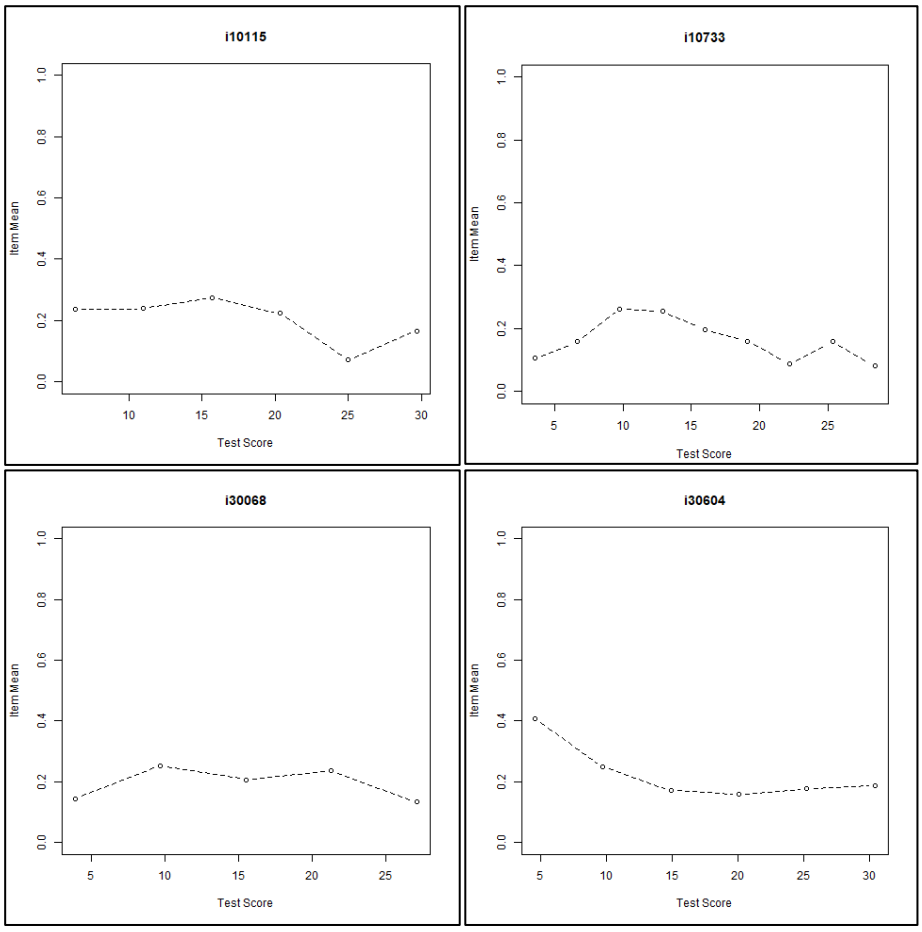


Figure 5: CTT Item Characteristic Curves – i10115 (English), i10733 (maths), i30068 (science) & i30604 (science)

A review of the actual test items revealed that each was poorly written and edited, with no clear correct answer. Two of the items are replicated in Figure 6 and it is plain to see why these items performed poorly, as the correct answers are not immediately obvious.

Item i10115, given to students age 11, asks the reader to identify a 'main problem' within a passage. The four options given as choices have no clear priority, with options b) and c) being indistinguishable. Arguably, option a) is about an activity rather than a problem, however the language in the question instructions also expects students to identify the conflict, which is then called the main problem, and option a) describes a conflict more than any other option. This item appears to have a poor stem, which should have been identified during item editing, and confusing distractors that have been highlighted by deeper data analysis. Item i10733, appears to be a straightforward knowledge recall question about leap years and the Wikipedia reasoning for a leap year (Multiple, 2020) explains the reasons for leap years as a mechanism to synchronise with seasonal years. The Wikipedia description matches most closely with option c), which is the correct answer, but the answer option is poorly written and factually incorrect. Again, this should have been identified during editing and the deeper data analysis during a pre-launch field test should have highlighted issues.

i10115	
Read the passage and identify the conflict (main problem) in it. Derek took several years to save money for his dream house. He finally bought one cottage near the seashore. Derek thought it was perfect! One fine day, he noticed a rat running around the house. First, there was only one, nibbling away at the food in the kitchen. Soon, there were two more. Derek had to deal with the rat menace. He went to war with the rats, one that he won, but at the cost of his dream house.	Choose your answer from the options. a) Derek fighting to end the rat menace b) Rats nibbling away at the food c) Derek noticing a rat running around the house. d) Derek taking several years to save money for a dream house. Correct = c

i10733	
The Earth takes 365¼ days to revolve around the Sun once. 365¼ days = 1 year Every four years, the four ¼ days add up to 1 day. This is added as an extra day in the fourth year. This year is called a leap year. Why is this done?	a) To add up a day in February b) To have the number of days in a year to be a whole number c) To have seasons in the same set of months every year d) To follow the Roman calendar Correct = c

Figure 6: Details of two test items

The worst performing items needed to be removed from the pool to improve the reliability each of the papers. Using the item review guidance method of Kline (2005), a group of items (24 in total) was eliminated and the CTT analysis was repeated. The Cronbach α for each paper was recalculated, and improvements in internal reliability were observed across most papers. In particular, the two least reliable papers, P-5-SCN-A1-Y18 and P-6-MATH-A1-Y-18, had improved reliability (α = 0.69 for both). This 'reduced set' of items was then used for further analysis.

4.5. Item Review – Item Response Theory

The IRT model parameters were estimated for each paper, using one-, two and three-parameter models with the mirt R package. The fit parameters for M2, root mean square error of approximation (RMSEA), comparative fit index (CFI) and Standardized Root Mean Square Residual (SRMR) were examined and in all cases the three-parameter model gave stronger fit statistics than one- or two-parameter models. A full set of item characteristic curves for the three-parameter model, paper by paper, was generated and inspected visually. Items with low or negative discriminations, or with high guessing parameters were eliminated, totalling 20 items, which now meant that 44 items overall had been removed – 20 from IRT based ICC inspections and 24 from CTT based ICC inspections. The Cronbach α for each paper was recalculated, and improvements in internal reliability were observed across most papers, and notably, all papers were now above an α of 0.70. With test reliability significantly enhanced by eliminating the poorly performing items, the data were evaluated for suitability for factor analysis (Kaiser, 1960).

4.6. Factor Analysis

The papers were evaluated through the KMO function in the R psych package and interpreted in relation to the Kaiser-Meyer-Olkin (KMO) Test. Three papers had a KMO index of factorial simplicity of just less than 0.80, but all other papers were > 0.80 or "meritorious" in Kaiser,

(1974) terms, and three papers were even “marvellous” (in the 0.90s). The Bartlett’s test of sphericity (Bartlett, 1950) was significant in all cases. The results, in Figure 7, indicated that the items were appropriate for factor analysis.

	P-5-ENG-A1-Y-18	P-5-ENG-E-Y-18	P-5-MATH-A1-Y-18	P-5-MATH-E-Y-18	P-5-SCN-A1-Y18	P-5-SCN-E-Y-18	P-6-ENG-A1-Y-18	P-6-ENG-E-Y-18	P-6-MATH-A1-Y-18	P-6-MATH-E-Y-18	P-6-SCN-A1-Y-18	P-6-SCN-E-Y-18
KMO	0.90	0.89	0.87	0.90	0.79	0.86	0.86	0.89	0.77	0.88	0.83	0.92
ChiSq	3839.6	3179.3	3254.1	2858.8	1896.6	2201.2	2995.4	4005.5	2573.9	4011.7	3103.9	5021.1
p	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

	P-7-ENG-A1-Y-18	P-7-ENG-E-Y-18	P-7-MATH-A1-Y-18	P-7-MATH-E-Y-18	P-7-SCN-A1-Y-18	P-7-SCN-E-Y-18	P-8-ENG-A1-Y-18	P-8-ENG-E-Y-18	P-8-MATH-A1-Y-18	P-8-MATH-E-Y-18	P-8-SCN-A1-Y-18	P-8-SCN-E-Y-18
KMO	0.79	0.84	0.83	0.84	0.88	0.90	0.86	0.86	0.86	0.87	0.80	0.89
ChiSq	2568.4	2199.2	3308.6	2261.5	4251.6	3152.7	3093.0	1984.2	3646.9	2237.4	2705.3	2197.0
p	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Figure 7: Kaiser-Meyer-Olkin Test and Bartlett’s test of sphericity with questionable items removed

An initial exploratory factor analysis was undertaken using the nScree function within the R nFactors package (Raiche, Gilles and David Magis, 2020), resulting in the summary in Figure 8. These results indicated a series of potential investigations and further exploratory work to reveal details of factors and loadings. The plot in Figure 9 highlights the nature of the factors for one paper and represents the eigenvalues as a scree plot (Cattell, 1966). This scree plot is representative of many papers in the assessment.

	P-5-ENG-A1-Y-18	P-5-ENG-E-Y-18	P-5-MATH-A1-Y-18	P-5-MATH-E-Y-18	P-5-SCN-A1-Y18	P-5-SCN-E-Y-18	P-6-ENG-A1-Y-18	P-6-ENG-E-Y-18	P-6-MATH-A1-Y-18	P-6-MATH-E-Y-18	P-6-SCN-A1-Y-18	P-6-SCN-E-Y-18	P-7-ENG-A1-Y-18	P-7-ENG-E-Y-18	P-7-MATH-A1-Y-18	P-7-MATH-E-Y-18	P-7-SCN-A1-Y-18	P-7-SCN-E-Y-18	P-8-ENG-A1-Y-18	P-8-ENG-E-Y-18	P-8-MATH-A1-Y-18	P-8-MATH-E-Y-18	P-8-SCN-A1-Y-18	P-8-SCN-E-Y-18
Optimal coordinates	2	2	2	3	3	1	3	3	3	3	3	3	4	2	3	3	4	2	3	2	4	1	3	1
Acceleration factor	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Parallel analysis	2	2	2	3	3	1	3	3	7	3	6	3	4	2	5	3	4	2	3	2	4	1	6	1
Kaiser rule	9	9	8	8	11	9	10	9	12	9	11	10	11	11	10	10	11	10	10	10	10	10	11	10

Figure 8: Factor analysis by paper

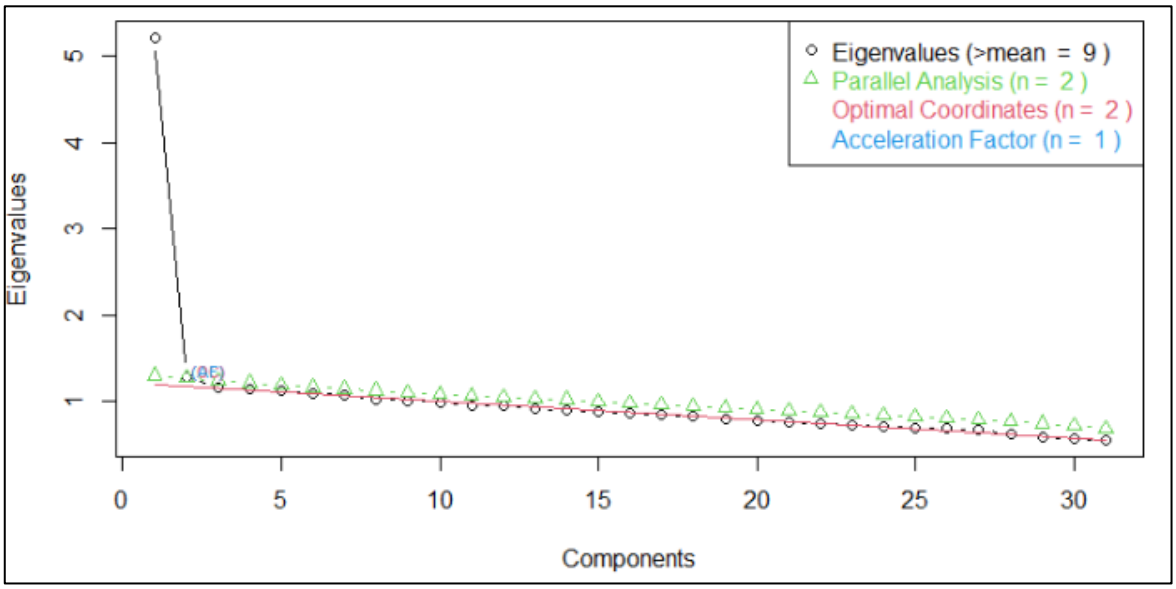


Figure 9: Scree plot of factors in paper P-5-ENG-A1-Y-18

Ahead of any further investigations, it is worth pausing to examine the factor summary in Figure 8 with reference to the intended competency and skills targets in the measurement items. In the original assessment design, Figure 2, the subjects papers target several competencies (5 for English, 4 for mathematics and 5 for science). This is at odds with the factor analysis (optimal coordinates) summary. The more limited factor analysis evaluation indicated that the test items failed to separate the competencies to the extent designed, and that more limited latent trait measurements were taking place. A factor analysis of the first paper (P-5-ENG-A1-Y18) was performed utilising the core R factanal function (R Core Team, 2020), which provides maximum-likelihood estimates, and a varimax rotation was reported.

Starting with 2 factors, as determined in table in Figure 8 for this paper, resulted in disappointing results. The item uniqueness was high across all items in the paper (> 0.69). The first factor carried a loading of 10.3% and the cumulative loading across the two factors was only 15.6%. However, the results showed the significance level of the χ^2 fit statistic is very small ($p << 0.0001$). This indicated that the hypothesis of perfect model fit is rejected. Progressively increasing the number of factors in the analysis demonstrated that the increase in cumulative factor loading was small for each new factor, but at 5 factors the fit statistic was no longer significant ($p = 0.0535$) – indicating that the hypothesis of perfect fit can longer be rejected. However, the cumulative factor loading was only 20.6%, leaving a considerable variance unexplained. As the number of factors was increased, the cross-factor loadings increased too, leading to a growing overlap or lack of separation between factors. Extending this analysis across all papers in the set yielded somewhat similar results. On some papers it was possible to specify factors to account for up to 30% of the variance, whilst revealing reasonable levels of loadings against those factors. In no case was it possible to reach acceptable variance explained proportions (Johnson & Wichern, 2007).

The factor analysis and extraction on the data paper-by-paper yielded disappointing results. However, the tests taken by children were designed to reveal a series of competency and skills that were expected to be demonstrated across subjects rather than within individual papers. The data were reprocessed to group results by child groupings and concatenate English, mathematics and science results into single sets, resulting in 8 groupings, 2 for each of the 4 year-groups, see Figure 10. Across the 8 groupings, there was good Cronbach's internal reliability ($\alpha > 0.882$).

Papers	Group
P-5-ENG-A1-Y-18	1
P-5-MATH-A1-Y-18	1
P-5-SCN-A1-Y18	1
P-5-ENG-E-Y-18	2
P-5-MATH-E-Y-18	2
P-5-SCN-E-Y-18	2
P-6-ENG-A1-Y-18	3
P-6-MATH-A1-Y-18	3
P-6-SCN-A1-Y-18	3
P-6-ENG-E-Y-18	4
P-6-MATH-E-Y-18	4
P-6-SCN-E-Y-18	4

Papers	Group
P-7-ENG-A1-Y-18	5
P-7-MATH-A1-Y-18	5
P-7-SCN-A1-Y-18	5
P-7-ENG-E-Y-18	6
P-7-MATH-E-Y-18	6
P-7-SCN-E-Y-18	6
P-8-ENG-A1-Y-18	7
P-8-MATH-A1-Y-18	7
P-8-SCN-A1-Y-18	7
P-8-ENG-E-Y-18	8
P-8-MATH-E-Y-18	8
P-8-SCN-E-Y-18	8

Figure 10: Children/Paper Groupings

Rerunning the nScree function across the groupings indicated the potential factor quantity to be retained, with results as shown in Figure 11.

	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7	Group 8
Optimal coordinates	6	5	6	5	4	2	9	5
Acceleration factor	1	1	1	1	1	1	1	1
Parallel analysis	8	5	9	5	11	5	9	5
Kaiser rule	37	36	39	35	38	35	38	37

Figure 11: Factor Analysis by Groups

Using the optimal coordinates values, as previously, exposed similar disappointing results – high levels of uniqueness, low cumulative factor loadings with large unexplained variances and some cross-loading. The assessment instrument in Group 1 consisted of 89 items with contexts that were a mixture of English, mathematics and science. Using a process of iterative removal of items (Clark & Watson, 1995), using a factor loading cut-off of 0.32 (Tabachnick & Fidell, 2001), a reduced item pool for Group 1 was derived, containing 18 items. The uniqueness for several items was lower than previously observed, and the 6 extracted factors explained 33% of the variance. The new instrument was not ideal, with a one factor loading against a single item, and one other factor against only two items. The comparison of these factors against the competencies, skills and sub-skills (Table 2) that had been intended in the assessment design showed alignment to subjects rather than to skills.

Factor	QuestionCode	SubjectName	Competency	Skill	SubSkill
1	i10068	English	Creative Thinking	Elaboration	Originality
1	i10079	English	Critical Thinking	Make Judgments	Synthesize information
1	i10086	English	Personal and social	Society	Protect and preserve environment
1	i10087	English	Personal and social	Interpersonal	Build and Manage relationships
2	i10694	Mathematics	Core Thinking	Acquisition	Memorization
2	i11368	Mathematics	Core Thinking	Application	Mathematical Fluency
2	i90114	Mathematics	Creative Thinking	Novelty	Explore possibilities
2	i90117	Mathematics	Core Thinking	Articulation	Information Organization
3	i10002	English	Core Thinking	Articulation	Exemplification
3	i10013	English	Core Thinking	Acquisition	Recognition and Assimilation
3	i10046	English	Communication	Present	Clarity
3	i10065	English	Creative Thinking	Novelty	Combine ideas
4	i10697	Mathematics	Core Thinking	Acquisition	Memorization
5	i90010	Science	Communication	Adapt	Observe
5	i90016	Science	Core Thinking	Acquisition	Recognition and Assimilation
5	i90181	Science	Creative Thinking	Novelty	Fluency in generating ideas
6	i11373	Mathematics	Core Thinking	Application	Mathematical Fluency
6	i11374	Mathematics	Core Thinking	Application	Mathematical Fluency

Table 2: Factor summary for reduced assessment instrument – Group 1

The above process could continue until a unidimensional measurement was achieved, but this assessment was never designed to measure a single construct, and specifically it was designed to measure a series of competencies so that formative feedback could be provided to children. A detailed look at each of the items, in the factor groupings revealed, might have enabled the identification of a latent trait that these factors were measuring against but this was considered as future research work. At this point, analysis of individual papers and groups of papers for year groups had been undertaken and a relatively limited number of latent traits were able to be reported against. However, it was necessary to construct a sub-set of the instruments to provide opportunity for cross-school analysis and for correlation with questions in the Psychometric paper. The target was to arrive at groups of questions that revealed a clear reporting of a set of latent traits that could be used across schools. The analysis of each of the groups was undertaken and sub-sets of items representing the strongest loading factors was created.

At this point, the analysis and research had focussed on Research question 1a:

What degree of construct validity exists in tests designed for diagnostic purposes?

How does a diagnostic assessment vary across schools, grades and subjects?

This has been done by first analysing a single paper and then a group of three papers, leaving opportunity for much further analysis of other papers and other groups of papers. This highlighted significant challenges and issues that provided indications of the assessment instrument capability, and what further research would be likely to expose.

It was considered appropriate to turn the analysis effort towards other aspects of the assessment and continue to explore the other research questions, and in particular, research question 1b:

> *What correlations in performance are apparent between cognitive and psycho-educational constructs?*

To this point the data analyses have focussed on the tests within subject contexts (English, mathematics and science). However, the students also complete a survey referred to as a Psychometric paper. This paper surveyed all students through a group of 40 questions, each with 4 choices that were marked from 0 to 4. This essentially provided a rating scale for each of the questions and generated an attribute degree (Linacre, 2002), although the degrees were ordinal, as each student option was judged to simply demonstrate more or less than another option (in the view of the question author) rather than to specifically illustrate an amount of difference. The questions are intended to measure various competencies and skills (see Figure 2 and Appendix B). The psych package within R was used to score and analyse the effectiveness of the psychometric questions (Revelle, 2013). The questions were grouped by their targeted skills and summarised, giving Cronbach α scores of:

```
      Intra Inter Adapt EmMgt Socty
alpha  0.33  0.39  0.27  0.43  0.53
```

The internal reliability of the questions was low and poorly performing items were suspected. The R software mirt was used to calculate item fit statistics and generate item characteristic curves for the polytomous items. This showed disappointing results, with many curves similar

to Figure 12, and the boundaries between categories were poorly defined and the discrimination was erratic.

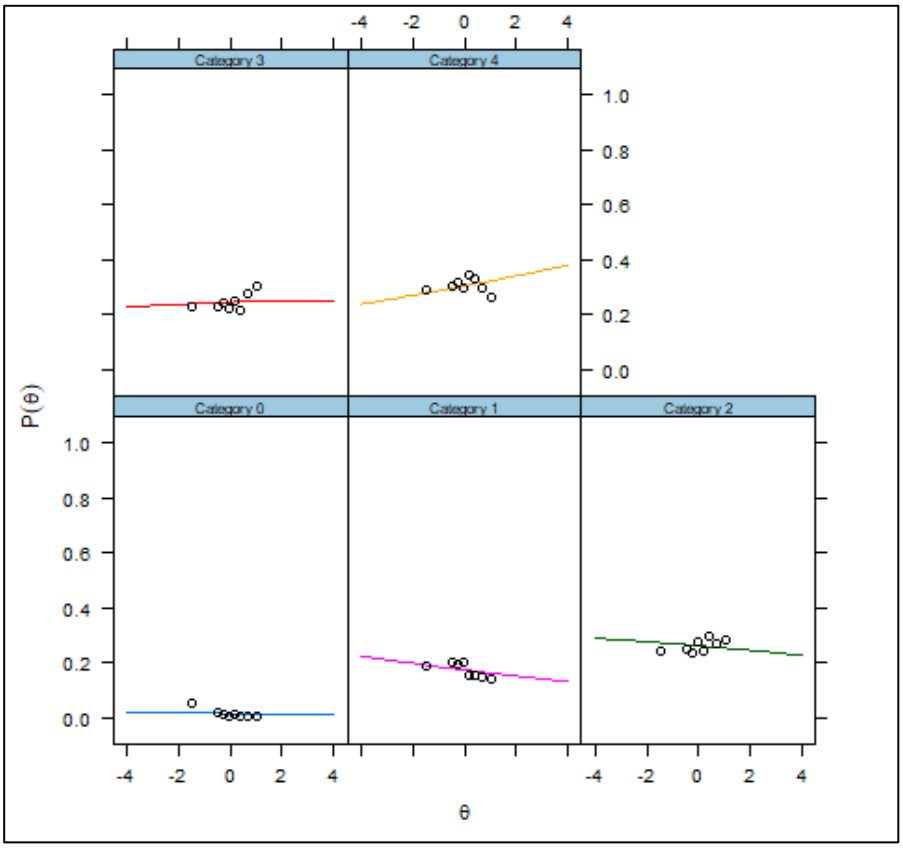


Figure 12: Item Characteristic curves for PSY-4

Items that performed poorly were removed for factor analysis to proceed. The nScree function reported the number of optimal coordinates as 4 and further factor analysis was undertaken. Variables were removed, using the Clark & Watson (2015) method, down to the 4 best loading variables. The p-value was not significant (p = 0.142), indicating that the hypothesis of perfect fit could no longer be rejected. The instrument was reduced to 11 questions that loaded most strongly against the 4 factors. The results are shown in Table 3 listing the items loading versus factors, and the total explained variance was 26%. Question PSY-32 loaded against 2 factors, although most strongly against Factor 4. Looking at the actual question in PSY-32 it is easy to see that the question was measuring 'intrapersonal' values and 'self-control' as per the factor analysis indication. Inspecting the individual questions and the intended skills helped to clarify potential category names for the factors.

At this stage, the psychometric paper had been reduced to 11 variables down from 40, measuring 4 skills, down from 5 skills with many sub-skills. The factor analysis demonstrated that there was significantly less discernment in the instrument than had been designed and that reporting against the many sub-skills was, in fact, unlikely to succeed. The overall variance explained in the reduced items was only 26%, which is generally viewed as unsatisfactory (Hair, Black, Babin, & Anderson, 2010, p. 108)

Factor	Skills	Items
1	Self-control	PSY-23, PSY-26, PSY-28, PSY-29, PSY-32
2	Interpersonal	PSY-15
3	Intrapersonal	PSY-32, PSY-33
4	Society	PSY-32, PSY-36, PSY-39, PSY-40

Table 3:Factor reduction of Psychometric paper

Considering student performance across subjects, we can see that there is a moderate positive correlation between the subjects (Table 4).

	ENG	*MAT*	*SCN*
ENG	1.00		
MAT	0.63	1.00	
SCN	0.63	0.68	1.00

Table 4: Subject to subject correlation

School to school comparisons showed that there was considerable variation. The mean score across all schools and all papers was 14.7 (SD = 6.42), and the range was 7.8 to 21.7. A correlation of mean scores by grade and by paper highlighted some moderate and strong positive correlations across the instruments (Table 5). Within grades, the correlations between subjects was high and greater than 0.82 in all cases. It would be expected that there would be high correlation between subject means of adjacent grades, which was generally the case, although the Grade 7 mathematics scores demonstrates a lower correlation with the Grade 8 mathematics scores (r = 0.75), than with the English (r = 0.81) and science (r = 0.85) scores.

This repeats for Grades 5 and 6 to a lesser extent and is possibly an indicator that the Grade 8 mathematics tests are problematic, or do not build on earlier years' knowledge.

		Grade 5			Grade 6			Grade 7			Grade 8		
		ENG	MAT	SCN	ENG	MAT	SCN	ENG	MAT	SCN	ENG	MAT	SCN
Grade 5	ENG	1.00											
	MAT	0.92	1.00										
	SCN	0.86	0.90	1.00									
Grade 6	ENG	0.92	0.87	0.90	1.00								
	MAT	0.86	0.89	0.85	0.91	1.00							
	SCN	0.79	0.80	0.86	0.88	0.90	1.00						
Grade 7	ENG	0.90	0.82	0.85	0.92	0.85	0.84	1.00					
	MAT	0.74	0.78	0.80	0.81	0.81	0.75	0.87	1.00				
	SCN	0.72	0.75	0.81	0.78	0.77	0.80	0.86	0.91	1.00			
Grade 8	ENG	0.80	0.81	0.86	0.86	0.85	0.79	0.86	0.81	0.80	1.00		
	MAT	0.71	0.81	0.82	0.74	0.82	0.71	0.71	0.75	0.72	0.82	1.00	
	SCN	0.76	0.82	0.90	0.83	0.83	0.84	0.86	0.85	0.89	0.93	0.85	1.00

Table 5: Correlation of school mean scores by grade and by paper

The data set only included anonymised school identifiers, so it was only possible to compare school to school performance numerically and categorically, with no contextualisation possible. It could have been illuminating to understand more about individual schools when exploring and explaining data variations and correlations.

Examining mean scores between schools highlights a generally even picture, except for school S017, which has a strong negative correlation with several schools. Schools S006 and S028 correlate weakly with many schools. Schools S002, S005 and S011 demonstrate the highest correlations across several schools. Without further contextual information for these schools, it would be impossible to investigate the reasons for the variable correlations. We could hypothesise many factors – teacher quality, curriculum mismatch, cohort differences, or any other number of factors – but the overriding sense is that the instruments will be sensitive to school related differences.

4.7. Correlations between cognitive and psycho-educational constructs

The data for all student total scores were initially used to explore scholastic (cognitive) to psychometric (psycho-educational) correlations. The psychometric instrument had been shown to be significantly deficient against the assessment framework underpinning the design, and factors in the data did not align with the intended measures of traits. The total

marks (ratings) for the psychometric results were used, although previously identified poor items were removed from the data. A series of linear regression studies were undertaken using the 'lm' function from the core R set of functions, beginning with simple regressions between the scholastic subject (as the dependent variables) and the psychometric scores (as the independent variant). The regression analysis was used to test if the psychometric scores significantly predicted student performance on the scholastic tests.

For English, the regression indicated that the psychometric test explained 7.3% of the variance (R^2 =0.073, p<0.0001). For maths, 7.6% of the variance was explained (R^2=0.076, p<0.0001), and for science the variance explained was 11.9% (R^2=0.119, p<0.0001). The linear regression plots for science and mathematics are shown in Figure 13 – the plot for English is not shown as it is almost identical to that for maths. The plot exposes significant bunching of the data, with outliers skewing the data. For science the regression, formula for the prediction of science (SCN) performance from psychometric (PSY) score is: SCN = 0.2 x PSY -0.5. Visually, without the outliers – those at the lower PSY ratings – the intercepts would be lower, and the slope would be significantly greater. The removal of outliers was beyond the scope of this research.

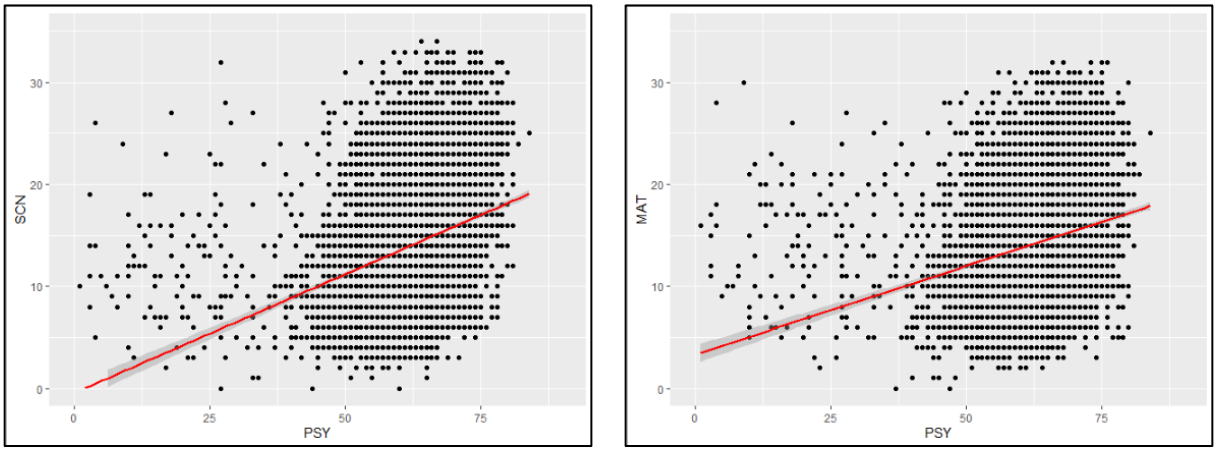


Figure 13: Linear Regression Plots for Science (SCN) and mathematics (MAT) versus psychometric (PSY)

4.8. Summary

The analysis explored a convenient, substantive data set using published methodologies to reveal the strength of construct validity in an assessment designed to measure competencies and skills in a country that is rooted in rote-learning. The assessment was designed to provide diagnostic information to students, parents, teachers and school leaders. Guidance from Herbert W. Marsh publications was used to guide the analysis, and in particular Marsh & Hau, (2007), which addressed definitions of construct validity and the methodologies for confirming validity. Research question 1 (What degree of construct validity exists in tests designed for diagnostic purposes?) was addressed by following the detailed stages outlined by Loevinger (1957) for test development: (i) formation of an item pool, (ii) analysis and selection of the final pool, and (iii) correlation of scores with criteria and variables. The DISCA item pool had been formed during the design of the assessment instrument (stage (i)) but further formation was explored. Significant analysis and selection were undertaken through reliability, CTT and IRT techniques (stage (ii)), and, finally, correlation, involving factor analysis, to create an optimum test design (stage (iii)).

The item analysis revealed significant issues driven by incorrect scoring rubrics, poor item editing and confusing distractor design. It had to be assumed that the instruments had not been sufficiently field trialled before formal use, as many of the identified issues would have surfaced in a field trial and the instrument could have been enhanced as a result. The analysis afforded an opportunity for some instrument enhancement to be performed and for internal reliabilities to be increased. Once a reduced and higher quality item pool subset had been identified, iterative factor analysis was done using exploratory factor analysis methodologies. The loadings of the emergent factors indicated significantly different trait measures to that intended in the design of the instrument. Also, the emergent factors explained a limited (only ~30%) of the variance and a much-reduced group of items could have been used to measure the traits.

Research question 1a (*How does a diagnostic assessment vary across schools, grades and subjects?)* was addressed in parallel with the analysis described under the three stage process; however, time limitation prevented analysis across all grades for all subjects. Representative papers, designed for specific ages, were examined individually by subject and grouped by grades to look for variations and correlations. The papers across grades and subjects indicated varying levels of reliability, and there was some indication that different papers or groups of papers would load against very different factors. This was demonstrated by the variable number of optimal coordinates, although it could be argued that the competencies and skill targeted at different grades led to the variability.

What correlations in performance are apparent between cognitive and psycho-educational constructs?

The data was used to explore scholastic (cognitive) to psychometric (psycho-educational) correlations. The psychometric instrument had been shown to be significantly deficient against the assessment framework underpinning the design, and factors in the data did not align with the intended measures of traits. The correlations appeared to be strongly skewed by a large population of outliers, the removal of which would probably have had a large impact on the slope and intercept parameters.

How useful is the Marsh and Hau Construct Validation Methodology for a diagnostic assessment where rote-learning dominates the school culture?

In simple terms, the methodology was shown to be effective as a guide in performing the research in this thesis. The challenge set by Marsh & Hau, (2007) – to close the gap between methodological researchers and substantive researchers – was taken and reported in this thesis. Whilst the methodology proved useful, the issue that emerged was that of another synergy gap – lack of synergy between researchers and implementors of assessments. The methodology highlighted issues with the assessment instruments, which must be seen as a

positive outcome as this opens an opportunity for quality improvement of the instruments. However, the instrument is a commercial assessment in use and implementors believe they have created and are using an effective assessment; the published research methodologies would say otherwise. The gap between researchers and implementors needs to close for the good of children in the education systems.

5. Discussion

This research set out to address a series of questions by using the construct validation methodology suggested by Marsh & Hau, (2007), who called for more methodological-substantive synergies to be investigated. Their methodology is clear, if rather high level and without specificity to the necessary tasks required. The methodology is appropriate for the exploration in this thesis, where we aim to understand challenges of designing new assessments in educational contexts that are more used to 'old fashioned' assessment approaches. The role of assessments should be to provide data and insight into student cohorts to help personalise and fine tune their learning (Agarwal, 2020; Shea & Duncan, 2013). To this end, assessments need to measure what skills children have and what they need to develop to improve, and this requires much more diagnostic or formative feedback, and certainly more than a simple grade or mark. The diagnostic information about children must report detail about skills, knowledge and abilities of children. The paper from the Council of the National Postsecondary Education Cooperative (2002) suggests that these (skills, knowledge and abilities) are founded on traits and characteristics of people, and are in themselves 'bundled' to define competencies. In a school setting, we understand that teachers need the diagnostic insights into their students with degrees of detail that allow them to plan a developmental roadmap for each child. If teachers need to know about student competencies, then these competencies must be described in terms of skills, abilities and knowledge, which in turn must be described in terms of traits and characteristics. The challenge in designing any diagnostic assessment is in creating instruments that provide precision at the right depth of measurement (AERA/NCME, 2014; Borsboom & Molenaar, 2015; Shea & Duncan, 2013). Ultimately the target for any assessment is defining constructs and then measuring against those, which takes us back to Marsh and Hau's work and the ambition for more researchers to use construct validation methodologies. As education

research fuels change in education systems there is a growing understanding of how assessment needs to evolve, to measure in a more effective way and to report more usefully, and how educationalists need to have better assessment literacy (DeLuca, LaPointe-McEwan, & Luhanga, 2016).

Education systems have assessment mechanisms, and many have historical traditions that influence present and future assessment programmes. The education systems that have historically relied on rote learning have assessment examples that target the measure of knowledge, where recall is the most successful strategy for a student, and assessments can do no more than report on a 'recall' competency. This leaves the teacher with little to go at from a developmental view, other than to teach to the test and resort to rote teaching to reinforce the memorisation process – a rather unworthy cycle (Black & Wiliam, 1998; Wiliam, 2000). Breaking away from this rote-recall cycle requires the teacher to be given information about students that is actionable, and for the teacher to realise developmental gains in their students. It also requires a reduction in teacher accountability (Clotfelter et al., 2004; Panesar-Aguilar, 2017) to deliver grades and marks in assessments, and a refocussing towards more personalised and targeted learning plans for students (Agarwal, 2020; Shea & Duncan, 2013). The research in this paper explored an attempt to define a diagnostic assessment in a country rooted in rote learning and examined challenges through analysis of secondary data captured in schools in India in August 2019.

The data available for the research in this thesis were large, necessarily, and therefore required specific and specialist applications for their handling. The R environment was chosen, in part due its flexibility and widespread support from package developers, in part due to the introductory teaching and learning provided by OUCEA under the MSc Educational Assessment and in part a desire by me, as the researcher and author, to use this as an extended learning experience. In retrospect, this was a good decision as it satisfied the learning objective whilst

enabling extensive, comprehensive, revealing and practically useful analysis. However, the R environment has many idiosyncrasies and the learning curve to reach a semi-proficient state is long (Chambers, 2014). At times many hours can be taken in researching the necessary approaches and functions of R (de Jonge & van der Loo, 2013). This eventually yields results and insights, but often provides only simple or single data points that in themselves cannot then be elaborated on very significantly. A skilled and experienced R user would work rapidly through early data analysis stages to reach deeper analyses quicker, allowing more significant discovery and investigation.

The data itself was available in a semi-structured form and was substantive and largely complete. A few issues were, however, there for discovery. Whilst detailed scores were available for all students, some scores were recorded with unexplained characters and it became apparent that only the 'attempted' results for students were provided. There was no access to the data collectors to audit or question the unexplained characters and the only option was removal of student results. The provision of only 'attempted' questions was less problematic, requiring a reasonable assumption that 'not attempted' equated to missing (Béland, Pichette, & Jolani, 2016; Chen et al., 2012). The questions appeared to be 'not attempted' in a random way, so it was fair to infer that students had not been able to address the question rather than not having been presented groups of questions.

Additionally, the questions were grouped into 'papers' and students were asked to complete full papers at one sitting. Finally, the recording of scores or scoring rubrics contained errors. This became apparent when significant analysis had been undertaken and one item was emerging as a strong indicator of a latent trait. In exploring the options for naming the trait and examining the question with allocated scores, it was obvious that high scores were given to wrong answers. It was beyond the time limits of this research to explore the scoring for each test item in this way, although that remains a necessary quality check for the assessment

overall. The impact of this issue was judged to be limited, except where analysis surfaced problems.

What degree of construct validity exists in a test designed for diagnostic purposes?

This research question was at the heart of the challenge and guidance introduced by Marsh and Hau. The India assessment is a carefully designed assessment based on an assessment framework created by experts in education. The framework elaborates competencies and skills that are intended to be measured across personality measures and scholastic ability. While the subject-based instruments are differentiated by age groupings, the personality (or psychometric) instrument (questionnaire) is used without modification across all ages. For the assessment to function as designed, the variables in the instruments are required to stimulate responses driven by the competencies of interest which then reveal information about the foundational traits of students. This requires the variables to be aligned to the traits they are measuring and to provide enough granularity and discrimination to allow reporting against a scale. The variables were used with students in 44 schools in August 2019 and across a total of over 7,000 students.

The assessment framework for the subject tests and the personality questionnaire was well conceived and robustly developed (Pearce et al., 2015), describing a competency-skill hierarchy consistent with the most current published research (Council of the National Postsecondary Education Cooperative, 2002). The commissioning of the items for the instruments was completed using publisher teams and locally based item writers, experienced at creating test items. However, this introduces the first questionable impact. Publishers are not assessment experts; they can review and edit content from a pedagogical view but not review assessment items from a measurement point of view. Additionally, any locally based item writers will be very experienced at writing in the contemporary style, which we know is oriented towards testing only knowledge (Burdett, 2016, 2017). It is very likely that

assessment items were created in an environment that built on knowledge test styles, reviewed by staff that knew little assessment theory or methodology. This combination reduced the possibility of creating assessment instruments that would measure the construct in the assessment framework. The results of the statistical analysis, the item characteristics review and iteration through factor analysis confirms that there were significant issues with construct validity. The data for the August 2019 tests were examined for general performance and reliability before exploring the factors and traits exposed through the data. This was compared to the intended constructs. Items were analysed using Classical Test Theory and Item Response Theory methods to reveal potential subsets of tests that could be better used for diagnosis of student skills.

Investigation of the 'scholastic' papers showed that the reliability statistics pointed to acceptable internal reliability, with some papers performing on at the borderline of acceptability. Inspection of item mean groupings versus total scores using classical test theory methodologies highlighted items that performed poorly. This was evident as either erratic discrimination where student total performance did not correlate with actual performance on certain items, and in some cases correlated negatively. Reviews of the questions and answer choices for the worst performing items revealed issues around lack of clarity of the questions' stems, unclear correct answers and overly enticing distractors.

This summary provides a harsh description of the assessments although removal of poor items was undertaken to improve the results and allow further analysis to continue. Item fit models were tested and evaluated, always leading to use of 3-parameter models for IRT analysis. This indicated that several items were being guessed at by students, and in some cases the guessing parameter was significantly higher than 25% (all items had 4 answer choices). The plots for the empirical data on item characteristics curves highlighted further item by item issues. Several items lacked monotonicity, and many demonstrated poor fit with high

residuals. Items were removed from the data and the combination of the CTT and IRT item culling resulted in severely reduced item pools, but with improved internal reliability of the instruments.

Factor analysis of the remaining data, with poorly performing items removed, was undertaken in an iterative methodology. The papers were evaluated through the Kaiser-Meyer-Olkin (KMO) Test (Kaiser, 1974), which confirmed the data were shown to be suitable for factor analysis (Kaiser, 1960) and nominal coordinate estimates were produced to guide the factorisation. Typically, the iteration through factor analysis, loadings reporting, item removal and back to factor analysis, resulted in a further reduced instrument with improved factor alignment. The resultant factors would typically explain only around 30% of the variance in the data, and adding factors did nothing to create better variance explanation when balancing against reasonable factor loadings. The assessments were designed to report on competencies and skills in the contexts of English, mathematics and science. Continuing the task set by Marsh and Hau, the analysis of the data collection against the constructs of interest began to reveal results. The *a priori* constructs for competencies and skills did not emerge strongly from the factor analysis, being largely overridden by the subject dimensions. For example, the Group 1 papers factorised into 6 factors of two English, three mathematics and one science, but with no overall alignment to the competencies and skills in the original construct.

The assessment package offered to schools is partly a ‘scholastic’ test and partly a ‘psychometric’ or personality test. The psychometric test was put through a similar analysis to that already discussed for the scholastic paper – reliability improvement through item performance investigation and iterative factor analysis. The Psychometric paper was designed around questions or statements of four option multiple choice that were scored in a graded way. This required generation of graded response item characteristic curves as groups of five, each modelling the measurement across grades. Although the ICCs were different to the

dichotomous ‘scholastic’ tests, the culling of items involved the same process of inspection and identified negatively discriminating items and/or items with poor residuals. The psychometric test was designed to be used across all ages, and the student results data was much greater (n = 7107) than for the individual scholastic papers. Iterative factor analysis identified a subset of items loading against 4 dominant factors. This was substantially different to the two competencies, five skills and 18 sub-skills intended in the assessment design.

The principal conclusion from this analysis must be that there was limited construct validity in relation to the competencies and skills targeted by the assessment but there was some measurement value in the instruments. However, the instruments could be significantly reduced in length to provide the same level of information as the full test.

How does a diagnostic assessment vary across schools, grades and subjects?

The data were provided with anonymised school references, for data privacy reasons, including only a simple school identifier code. The data did not include any school characteristics so any statistical analysis in relation to the schools was purely on a categorical basis. This did highlight that a few schools correlated poorly or negatively with the full school population, and that only a few schools correlated more strongly with others. In general, we can deduce that a good deal of variability existed in the schools. However, we can only hypothesise about these differences using literature guidance (O’Dwyer, 2005); but, the overriding sense is that the instruments will be sensitive to school related differences. This is not presented as a negative aspect of the DISCA instrument as diagnostic measures will be influenced by teacher quality, curriculum mismatch, cohort differences, and many other factors. The instrument rightly needs to be sensitive to these differences and the reporting to students needs to reveal those aspects as contributors to student performance.

What correlations in performance are apparent between cognitive and psycho-educational constructs?

The data included cross-age, self-reported contextual and personality data, called Psychometric in the instrument design. This was examined for correlations and dependencies across cognitive and non-cognitive aspects. The psychometric tests included slightly more missing data than the scholastic tests, but still within an acceptable level. It is tempting to hypothesise, based on literature guidance (Newman, 2014), about the reasons for the higher level of missing data and there is potential for this to be related to the context in which the instrument was used – in this case, a country more used to tests for a rote-learning environment where psycho-educational tests are less understood and less practiced. A substantive research study should be designed to prove or disprove this hypothesis.

The psychometric test data were slightly skewed and leptokurtic. The item by item analysis produced a large list of problematic items, with a variety of issues being revealed. One item was initially emerging from the factor analysis as a strong indicator of a latent trait and in trying to name the item's trait, it became apparent that the item had been coded in reverse to its correct rating. This was very likely to be an issue with the item's rubric, although there is no specific evidence either way. We could hypothesise based on literature (Brookhart & Chen, 2015), but this is a challenge for future work. The item was removed from any further analysis, and it was beyond the scope of this research to review each item for this type of issue. The psychometric tests were intended to measure five main competencies through items addressing 18 skills.

The factor analysis of the data found that possibly four latent traits could be discerned (Self-control, Interpersonal, Intrapersonal, Society). The item pool could be reduced from 40 items to just 12, although 17 of the items (nearly half of the paper) required removal due to their poor measurement and discrimination. The investigation into the predictive strength of the

different constructs highlighted that there were many outliers that influenced the regression formulae. Removing outliers was beyond the scope of this research but would be a necessary activity to enhance the deeper investigation into the performance of the DISCA instrument.

In the context of this research some of the challenges that emerge from the correlation of the cognitive and the psycho-educational constructs can be summarised as falling into three categories: the process for editing and review of the rubrics is poor; a cycle of reduction or refinement of the size of the item pool was missing; and the reasons for varying levels of missing data needs to investigated.

How useful is the Construct Validation Methodology for a diagnostic assessment where rote-learning dominates the school culture?

Marsh and Hau (2007) propose that the Construct Validation Methodology should be employed more frequently in substantive research. In using the methodology for review of the data for this thesis, we understood its applicability in this specific context (country and assessment type). It was shown to be highly effective at guiding the approach and provided relevant links to other literature. There were parts of the full methodology that were not reached in the research in this thesis (for example, structural equation modelling), so future research will continue to implement the methodology. This thesis brought together a series of methodological approaches with substantive data, and, although not all aspects of their methodology were reached, it was ultimately confirmed that the approach and methodology was useful and effective.

6. Conclusions and Recommendations

6.1. Conclusions

This thesis explores the challenges of developing diagnostic assessments in countries rooted in rote learning, by analysing an instrument designed in India for schools and students in the country. The research uses a convenient data set from a test session in August 2019 within 44 schools. The school system in India is very much rooted in a culture of rote learning, where grades are everything and teaching to the test is very common. This environment is not conducive to helping children identify their best competencies and skills, even though there is recognition that this would help to enhance a child's learning. The DISCA assessment from ABC Ltd offers insights into the skills and competencies of children, in a way that could provide a tuned or enhanced teaching and learning personalisation.

The ambition for the instrument is worthy, but it must deliver on the promise made. The first main research question in this thesis asks: *What degree of construct validity exists in tests designed for diagnostic purposes?* If the tests do not measure the competencies and skills being targeted, then the reporting to schools, children and their parents would be flawed.

The work of Marsh & Hau (2007) was chosen as a guiding methodology for evaluating construct validity as these authors are very significant contributors in this field and their work is much referenced. Furthermore, they set the challenge for researchers to bring together methodological approaches with substantive data to seek out synergies. The use of this methodology was the focus of the second main research question in this thesis: *How useful is the Marsh and Hau Construct Validation Methodology for a diagnostic assessment where rote-learning dominates the school culture?*

Although the instrument used to generate the data for the research had been through a benchmarking (standards setting) cycle and piloting before formal use, it was prudent to

recheck the performance for test items and the reliability of test papers. The three-step approach laid out by Loevinger (1957) was followed: formation of an item pool, analysis and selection of the final pool, and correlation of scores with criteria and variables.

Several items were found to be deficient, which is, perhaps, surprising when considering the level of development, benchmarking and pilots that had taken place. There was evidence of poor question stem authoring, rubric errors and confusing distractor design. It must be concluded that the test item benchmarking and piloting was not sufficient to generate a robust test instrument.

Poorly performing items were removed from the data set to enhance overall test reliability. Following the experience of the item level analysis it was decided to use exploratory rather than confirmatory factor analysis. This was performed using individual test papers and groups of test papers and comparisons of the emergent factors were made against the original intended competencies and skills. It was possible to discern some alignment, but at a relatively superficial level. The promise of measuring competencies and skills could not be fully supported by the research in this thesis.

This data analysis signals a poor construct validity. There is no doubt that issues with the test items were exposed, but other conclusions can only be hypothesised and investigating those hypotheses would require carefully designed research projects. It can be hypothesised that students were not used to taking tests of this nature, especially as there is no chance for a student to become accustomed to them ahead of formal testing. The tests have only been in existence since 2017 and only used in a small number of sessions, which implies that individual children have probably only been involved in these tests in a limited way. In a country where rote type tests are prevalent students might be surprised by competency and skills tests. As the instrument is used more over time, students will become more accustomed to these types of tests and might approach them in a different way. It could also be hypothesised that the

assessment industry, in a country where rote learning is prevalent, does not yet have the expertise to create instruments of this type. Although the assessment framework guiding this instrument is clear, the item writing, editing, benchmarking, piloting and publishing processes have not converted the framework into a fully effective instrument.

Returning to the second main research question – *How useful is the Marsh and Hau Construct Validation Methodology for a diagnostic assessment where rote-learning dominates the school culture?* – it must be concluded that the methodology is effective. The methodology exposed deficiencies of construct validity in an assessment instrument that had been through a supposedly thorough development process. The construct validation methodology was a valuable signpost towards other methodologies and should continue to be used by researchers when developing assessments. As Marsh and Hau (2007) point out, there is a disconnect between methodological researchers and substantive researchers, partly catalysed by the assessment community. The other disconnect, or 'gap', that needs to recognised is that between implementors of assessments and the research community.

The methodology is described in a scientific paper, making it rather inaccessible to those practically implementing the work in the methodology. It can be argued that implementers of these types of methodologies should be able to access this type of publication, but it could be equally argued that there is a need for this type of publication to be made more accessible. Potentially, there are not enough highly skilled assessment implementors, able to access this literature, and, potentially, authors are not writing in a way that reaches the implementors. This 'gap' between the two is a void that could be filled from both sides – with more accessible literature and better trained implementors. This is precisely the point that Marsh and Hau argue for – methodological-substantive synergies – although they argued a lack of synergy between researchers and implementors, rather than between methodological researchers and substantive researchers.

6.2. Recommendations for Commercial Producers of Instruments and Exam Boards

The conclusions in this thesis point to a series of recommendations for those designing this form of instrument in this context. These recommendations, founded on published materials, are summarised here and advice for further literature reviewing is given. The recommendations are given to commercial producers and exam boards together, as these two need to collaborate and strive for the same principles and standards for the good of the children in their systems.

The guidance from Loevinger (1957) must be a starting point and the three-step process should be adopted: formation of an item pool; analysis and selection of the final pool; and correlation of scores with criteria and variables. The third step links into the work of Marsh & Hau (2007), where an overarching methodology is described and many processes within that methodology are signposted. The work of Clark & Watson (1995) is particularly helpful in providing practical guidance for achieving construct validity when developing scales.

This thesis has shown that these methodologies and processes are illuminating and have practical applications. There were issues within the instrument at the centre of this thesis that need further investigation and, ultimately, remediation. ABC Ltd is encouraged to create a full methodology for converting the assessment framework into an edited, reviewed, trialled and refined instrument, based on processes in the published work referenced here.

Additionally, further literature reviewing (Anderson, Kellogg, & Gerbing, 1988; T. A. Brown, 2006; Marsh, Hau, et al., 2009; Rosseel, 2012; Tabachnick & Fidell, 2001) and empirical application would support the refinement of these processes and the enhancement of the overall methodology efficacy.

6.3. Recommendations for Further Research

The assumption of the researcher in entering this research project was that the DISCA instrument had been through full and thorough design and development. Many forms of investigation and analysis were carried out during the research for this thesis and a large amount of instrument refinement was necessary to enable more extensive analysis. Given the limit of time for research for this thesis many avenues for further research remained open and these are recommended here.

The data contained an acceptable level of missing responses which appeared to be random and were assumed in the research to be Missing Completely at Random (MCAR), although also appeared to be very variable. One investigation should examine if missing data were more prevalent at the end of the papers and were due to time restrictions for students. This would require the available item sequence data to be correlated with the missing data. Another investigation should examine any grouping of missing data around items or students or schools to reveal item design issues or contextual factors influencing lack of responses. The investigation into missing data should be linked to research into the performance of distractors in the test papers. This would be equally relevant to scholastic test and psychometric tests. The psychometric tests were of a rating scale design, but around a four-option multiple-choice format, meaning that each response was more categorical than interval. Many items were removed from the data and some of these items could have overly confusing distractors. Research into the distractors could potentially provide recommendations as to how the items could be edited and enhanced.

Examination of item characteristic curves led to the removal of many items from the data. This filtering of items was performed substantially through visual inspection of the curves under guidance of Kline (2005), but this area would benefit from development of a set of criteria to apply as a filter. The visual inspection criteria were used to identify items with: overly large

guessing parameters; negative or very low discriminations; and high residuals versus the calculated curve. A further research study could precisely enumerate these filters and apply them to identifying poor items. A potential approach would be to use the mirt mod2values function (Chalmers, 2012), which converts the mirt models into tables of parameters that can then be more numerically analysed.

The item pool was used in varying groups across papers, with some papers containing items used in other papers and some items appearing uniquely in single papers. Combining common items from different papers would add statistical strength to some of the analyses and this, too, should form part of further research studies. This may improve the factor analysis results and clarify the potential for discernment of latent traits for improvement of reporting. When creating larger groups of items, the process of item pool selection could be repeated which may expand the useable item pool or may indeed further restrict it. It would be interesting to consider the attenuation paradox (Loevinger, 1954) in this iterative reduction of the item pool and determine the absolute best balance.

6.4. Research Limitations

Very little characteristic data in relation to schools was included in the data set and this limited the school-to-school analysis. It is recommended that the data capture should enhance this by including such characteristics as: school size, school league table rating, gender and ethnicity factors, schools' location, funding parameters, and any other potentially useful information. More extensive contextual data would enable differential item functioning (Andrich & Hagquist, 2015; Kreiner & Christensen, 2014; Magis, Béland, Tuerlinckx, & de Boeck, 2010) and multi-level modelling techniques (T. A. Brown, 2006; Tabachnick & Fidell, 2001) to be utilised. This in turn would help to provide more extensive insight into the contextual dependencies of the assessment instruments.

The regression analysis pointed towards a significant group of outliers in the data and a new research strand around these is recommended. The regression plots signalled correlation between psychometric and scholastic performance but with a possible masking of the correct slope and intercept parameters due to outliers. Understanding the relationship between psychometric and scholastic results would help to tune the quality of the assessment instrument.

The setting for the research data was India, where rote learning dominates and where tests are designed for rote learning environments. Potentially, the tests that students faced were so different to the type of test that they were used to that there was a lack of content validity. This tension between content validity and construct validity should be explored.

7. Bibliography

AERA/NCME. (2014). *Standards for educational and psychological testing*.

Agarwal, P. (2020). RECENT TRENDS OF SMART CLASSES AND digitalization of education in India. *UGC Care Journal*, *31*(08), 703–718.

Anderson, J. C., Kellogg, J. L., & Gerbing, D. W. (1988). Structural Equation Modeling in Practice: A Review and Recommended Two-Step Approach. In *Psychological Bulletin* (Vol. 103).

Andrich, D., & Hagquist, C. (2015). Real and Artificial Differential Item Functioning in Polytomous Items. *Educational and Psychological Measurement*, *75*(2), 185–207. https://doi.org/10.1177/0013164414534258

Areepattamannil, S. (2014). Relationship between academic motivation and mathematics achievement among indian adolescents in Canada and India. *Journal of General Psychology*, *141*(3), 247–262. https://doi.org/10.1080/00221309.2014.897929

Baird, J.-A., Isaacs, T., Johnson, S., Stobart, G., Yu, G., Sprague, T., & Daugherty, R. (2011). *POLICY EFFECTS OF PISA*.

Banchariya, S. (2019). What could PISA 2021 mean for India - Times of India. Retrieved July 22, 2020, from The Times of India website: https://timesofindia.indiatimes.com/home/education/news/what-could-pisa-2021-mean-for-india/articleshow/67835819.cms

Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Psychology*, *3*, 77–85.

Béland, S., Pichette, F., & Jolani, S. (2016). *Impact on Cronbach ' s α of simple treatment methods for missing data*. *12*(1), 57–73.

Black, P., & Wiliam, D. (1998). Inside the Black Box: Raising Standards Through Classroom Assessment. In *Numismatic Chronicle* (Vol. 177).

Borsboom, D., & Molenaar, D. (2015). Psychometrics. In *International Encyclopedia of the Social & Behavioral Sciences* (pp. 418–422). https://doi.org/10.1016/B978-0-08-097086-8.43079-5

Brookhart, S. M., & Chen, F. (2015). The quality and effectiveness of descriptive rubrics. *Educational Review*, *67*(3), 343–368. https://doi.org/10.1080/00131911.2014.929565

Brown, G. T. L. (2011). School based assessment methods: Development and implementation. *Journal of Assessment Paradigms*, 30–32.

Brown, T. A. (2006). Methodology in the Social Sciences. In *Methodology in the Social Sciences*.

Browne, E. (2016). *Evidence on formative classroom assessment for learning Question A literature review on research, evidence and programmatic approaches on formative classroom assessment for learning*.

Burdett, N. (2016). *The good, the bad, and the ugly – testing as a part of the education ecosystem*.

Burdett, N. (2017). *Review of High Stakes Examination Instruments in Primary and Secondary School in Developing Countries*. (December), 1–55. Retrieved from www.riseprogramme.org

Byrne, B. M. (2012). *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. routledge.

Cai, L., & Hansen, M. (2013). Limited-information goodness-of-fit testing of hierarchical item factor models. *British Journal of Mathematical and Statistical Psychology*, *66*(2), 245–276. https://doi.org/10.1111/j.2044-8317.2012.02050.x

Caro, D. H., Sandoval-Hernández, A., & Lüdtke, O. (2014). Cultural, social, and economic capital constructs in international assessments: an evaluation using exploratory structural equation modeling. *School Effectiveness and School Improvement*, *25*(3), 433–450. https://doi.org/10.1080/09243453.2013.812568

Cattell, R. B. (1966). The Scree Test For The Number Of Factors. *Multivariate Behavioral Research*, *1*(2), 245–276. https://doi.org/10.1207/s15327906mbr0102_10

Chalmers, P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/doi:10.18637/jss.v048.i06

Chambers, J. M. (2014). Object-oriented programming, functional programming and R. *Statistical Science*, *29*(2), 167–180. https://doi.org/10.1214/13-STS452

Chen, S. F., Wang, S., & Chen, C. Y. (2012). A simulation study using EFA and CFA programs based the impact of missing data on test dimensionality. *Expert Systems with Applications*, *39*(4), 4026–4031. https://doi.org/10.1016/j.eswa.2011.09.085

Child, D. (1990). *The essentials of factor analysis.* Cassell Educational.

Choi, Y. J., & Asilkalkan, A. (2019). R Packages for Item Response Theory Analysis: Descriptions and Features. *Measurement*, *17*(3), 168–175. https://doi.org/10.1080/15366367.2019.1586404

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, *7*(3), 309–319. https://doi.org/10.1037/14805-012

Clarke, M. (2012). What Matters Most for Student Assessment Systems: A Framework Paper. *READ/SABER Working Paper Series. …*, *1*, 40. Retrieved from http://www-wds.worldbank.org/external/default/WDSContentServer/WDSP/IB/2012/04/24/000386194_20120424010525/Rendered/PDF/682350WP00PUBL0WP10READ0web04019012.pdf

Clotfelter, C. T., Ladd, H. F., Vigdor, J. L., & Diaz, R. A. (2004). Do School Accountability Systems Make It More Difficult for Low-Performing Schools to Attract and Retain High-Quality Teachers? *Journal of Policy Analysis and Management*, *23*(2), 251–271. https://doi.org/10.1002/pam.20003

Colliver, J. A., Conlee, M. J., & Verhulst, S. J. (2012). From test validity to construct validity … and back? *Medical Education*, *46*(4), 366–371. https://doi.org/10.1111/j.1365-2923.2011.04194.x

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research and Evaluation*, *10*(7).

Council of the National Postsecondary Education Cooperative. (2002). Report of the national postsecondary education cooperative working group on competency-based in postsecondary education. *U.S. Department of Education, National Center for Education Statistics*, 7–9.

de Jonge, E., & van der Loo, M. (2013). An introduction to data cleaning with R. In *R package*. https://doi.org/60083 201313- X-10-13

Deb, S. (2018). Learning and Education in India: Social Inequality in the States. *Social Change*, *48*(4), 630–633. https://doi.org/10.1177/0049085718802533

DeLuca, C., LaPointe-McEwan, D., & Luhanga, U. (2016). Teacher assessment literacy: a review of international standards and measures. *Educational Assessment, Evaluation and Accountability*, *28*(3), 251–272. https://doi.org/10.1007/s11092-015-9233-6

Dyer, N. G., Hanges, P. J., & Hall, R. J. (2005). Applying multilevel confirmatory factor analysis techniques to the study of leadership. *Leadership Quarterly*, *16*(1), 149–167. https://doi.org/10.1016/j.leaqua.2004.09.009

Ferrer, E., & McArdle, J. J. (2003). The Best of Both Worlds: Factor Analysis of Dichotomous Data Using Item Response Theory and Structural Equation Modeling. *Structural Equation Modeling*, *10*(4), 493–524. https://doi.org/10.1207/s15328007sem1004

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct Validation in Social and Personality Research: Current Practice and Recommendations. *Social Psychological and Personality Science*, *8*(4), 370–378. https://doi.org/10.1177/1948550617693063

Fox, J. (1987). Effect Displays for Generalized Linear Models. *Sociological Methodology*, *17*, 347. https://doi.org/10.2307/271037

Furr, R. M. (2018). *Psychometrics: An Introduction* (Third). London: Sage.

Google. (2020). Google Scholar Search. Retrieved July 25, 2020, from https://scholar.google.com/schhp?hl=en&as_sdt=0,5

Google Scholar. (2020a). Hau Kit-Tai - Google Scholar. Retrieved July 25, 2020, from https://scholar.google.com/citations?user=G-c_YRAAAAAJ&hl=en&oi=sra

Google Scholar. (2020b). Herbert W. Marsh - Google Scholar. Retrieved July 25, 2020, from https://scholar.google.com/citations?hl=en&user=w911YWwAAAAJ&view_op=list_works

Gove, M. (2014). Time to tear down the walls.

Graham, J. W., & Hofer, S. M. (2000). Multiple imputation in multivariate research. *Modeling Longitudinal and Multilevel Data:Practical Issues, Applied Approaches, and Specific Examples*, (2000), 201–218.

Guo, J., Marsh, H. W., Parker, P. D., Dicke, T., Lüdtke, O., & Diallo, T. M. O. (2019). A Systematic Evaluation and Comparison Between Exploratory Structural Equation Modeling and Bayesian Structural Equation Modeling. *Structural Equation Modeling*, *26*(4), 529–556. https://doi.org/10.1080/10705511.2018.1554999

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate Data Analysis*.

Jennings, J. (2012). The effects of accountability system design on teachers' use of test score data. *Teachers College Record*, *114*(11), 1–23.

Johnson, R. A., & Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis* (Sixth). Pearson Education Inc.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Measurement*, *XX*(1), 141–151.

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, *39*(1), 31–36. https://doi.org/10.1007/BF02291575

Kline, R. B. (2015). *Principles and Practice of Structural Equation Modeling* (3rd ed.). https://doi.org/10.5840/thought194520147

Kline, T. (2005). *Psychological Testing: A Practical Approach to Design and Evaluation*. https://doi.org/10.4135/9781483385693

Kreiner, S., & Christensen, K. B. (2014). Analyses of Model Fit and Robustness. a New Look At the Pisa Scaling Model Underlying Ranking of Countries According To Reading Literacy. *Psychometrika*, 210–231.

Liem, G. A. D., & Martin, A. J. (2013). LATENT VARIABLE MODELING IN EDUCATIONAL PSYCHOLOGY: INSIGHTS FROM A MOTIVATION AND ENGAGEMENT RESEARCH PROGRAM. In *Application of structural equation modeling in educational research and practice* (pp. 187–216). Sense Publishers.

Linacre, J. (2002). Understanding Rasch measurement: Optimizing Rating Scale Category Effectiveness. *Journal of Applied Measurement*, *3*, 85–106.

Little, R. J. A., & Rubin, D. B. (2002). *Introduction 1.1.Statistical Analysis with Missing Data*. John Wiley & Sons, Incorporated.

Loevinger, J. (1954). The attenuation paradox in test theory. *Psychological Bulletin*, *51*(5), 493–504. https://doi.org/10.1037/h0058543

Loevinger, J. (1957). OBJECTIVE TESTS AS INSTRUMENTS OF PSYCHOLOGICAL THEORY1. In *Psychological Reports* (Vol. 3). @ Southern Universities Press.

Magis, D., Béland, S., Tuerlinckx, F., & de Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods*, *42*(3), 847–862. https://doi.org/10.3758/BRM.42.3.847

Marsh, H. W., Hau, K.-T., & Wen, Z. (2009). *Structural Equation Modeling In Search of Golden Rules: Comment on Hypothesis-Testing Approaches to Setting Cutoff Values for Fit Indexes and Dangers in Overgeneralizing Hu and Bentler's (1999) Findings*. https://doi.org/10.1207/s15328007sem1103_2

Marsh, H. W., & Hau, K. T. (2007). Applications of latent-variable models in educational psychology: The need for methodological-substantive synergies. *Contemporary Educational Psychology*, *32*(1), 151–170. https://doi.org/10.1016/j.cedpsych.2006.10.008

Marsh, H. W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J. S., & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. In *Structural Equation Modeling* (Vol. 16). https://doi.org/10.1080/10705510903008220

Miri, B., David, B. C., & Uri, Z. (2007). Purposely teaching for the promotion of higher-order thinking skills: A case of critical thinking. *Research in Science Education*, *37*(4), 353–369. https://doi.org/10.1007/s11165-006-9029-2

Multiple. (2020). Leap year - Wikipedia. Retrieved July 26, 2020, from https://en.wikipedia.org/wiki/Leap_year

National Research Council. (1996). Cap 1- Science Content Standards. In *National Science Education Standards*. https://doi.org/10.17226/4962

Newman, D. A. (2014). Missing Data: Five Practical Guidelines. *Organizational Research Methods*, *17*(4), 372–411. https://doi.org/10.1177/1094428114548590

O'Dwyer, L. M. (2005). Examining the variability of mathematics performance and its correlates using data from TIMSS '95 and TIMSS '99. *Educational Research and Evaluation*, *11*(2), 155–177. https://doi.org/10.1080/13803610500110802

OECD. (2017). *PISA 2015 Assessment and Analytical Framework*. https://doi.org/10.1787/9789264281820-en

OECD. (2019). Germany's PISA Shock - OECD. Retrieved July 22, 2020, from https://www.oecd.org/about/impact/germanyspisashock.htm

Orcan, F. (2018). Exploratory and Confirmatory Factor Analysis: Which One to Use First? *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, (February), 414–421. https://doi.org/10.21031/epod.394323

Panesar-Aguilar, S. E. (2017). Promoting Effective Assessment for Learning Methods to Increase Student Motivation in Schools in India. *Research in Higher Education Journal*, *32*, 1–16.

Pearce, J., Edwards, D., Fraillon, J., Coates, H., Canny, B. J., & Wilkinson, D. (2015). The rationale for and use of assessment frameworks: improving assessment and reporting quality in medical education. *Perspectives on Medical Education*, *4*(3), 110–118. https://doi.org/10.1007/s40037-015-0182-z

Popham, W. J. (2011). Assessment literacy overlooked: A teacher educator's confession. *Teacher Educator*, *46*(4), 265–273. https://doi.org/10.1080/08878730.2011.605048

Qadir, J., Taha, A. M., Yau, K. A., Ponciano, J., Hussain, S., Al-fuqaha, A., & Imran, M. A. (2020). *Leveraging the Force of Formative Assessment & Feedback for Effective Engineering Education*. 1–23.

R Core Team. (2020). *R: A language and environment for statistical computing*. Retrieved from https://www.r-project.org/

Raiche, Gilles and David Magis, D. (2020). *nFactors: Parallel Analysis and Other Non Graphical Solutions to the Cattell Scree Test.*

Raîche, G., Riopel, M., & Blais, J.-G. (2006). Non graphical solutions for the Cattell's scree test. *International Meeting of the Psychometric Society*, 1–12.

Raîche, G., Walls, T. A., Magis, D., Riopel, M., & Blais, J. G. (2013). Non-graphical solutions for Cattell's scree test. *Methodology*, *9*(1), 23–29. https://doi.org/10.1027/1614-2241/a000051

Revelle, W. (2013). *Using R to score personality scales ∗*. 1–11. Retrieved from http://personality-project.org/r/psych/HowTo/scoring.pdf

Revelle, W. (2018). How to: Use the psych package for factor analysis and data reduction. *Rdrr.Io*, 1–86. Retrieved from https://rdrr.io/cran/psychTools/f/inst/doc/factor.pdf

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2).

Ruscio, J., & Roche, B. (2012). Determining the number of factors to retain in an exploratory factor analysis using comparison data of known factorial structure. *Psychological Assessment*, *24*(2), 282–292. https://doi.org/10.1037/a0025697

Schauberger, P., & Walker, A. (2020). *openxlsx: Read, Write and Edit xlsx Files*.

Shea, N. A., & Duncan, R. G. (2013). From Theory to Data: The Process of Refining Learning Progressions. *Journal of the Learning Sciences*, *22*(1), 7–32. https://doi.org/10.1080/10508406.2012.691924

Singh, A. (2020). India To Participate In PISA 2021. Know What Is PISA. Retrieved July 22, 2020, from NDTV Education website: https://www.ndtv.com/education/india-to-participate-in-pisa-2020-know-what-is-pisa-2177883

Soland, J., Hamilton, L. S., & Stecher, B. M. (2013). Measuring 21st century competencies: Guidance for educators. *Asia Society Global Cities Education Network Report*, (November), 68. Retrieved from http://asiasociety.org/files/gcen-measuring21cskills.pdf

Spearman, C. (1904). " General Intelligence ," Objectively Determined and Measured Author (s): C . Spearman Source : The American Journal of Psychology , Vol . 15 , No . 2 (Apr ., 1904), pp . 201-292 Published by : University of Illinois Press Stable URL : http://www.jsto. *The American Journal of Psychology*, *15*(2), 201–292.

Stambach, A., & Hall, K. (2016). *Anthropological perspectives on student futures : youth and the politics of possibility*.

Stobart, G. (2008). Testing times: The uses and abuses of assessment. *Testing Times: The Uses and Abuses of Assessment*, 1–218. https://doi.org/10.4324/9780203930502

Tabachnick, B. G., & Fidell, L. S. (2001). Multivariate Statistics. In *Using Multivariate Statistics*. https://doi.org/10.1007/978-1-4757-2514-8_3

Tucker, L., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*(1), 1–10.

Warsi, L. Q., & Shah, A. F. (2019). Teachers' perception of Classroom Assessment Techniques (CATs) at Higher Education Level. In *Pakistan Journal of Social Sciences (PJSS)* (Vol. 39).

Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software*, *21*(12).

Wickham, Hadley, Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wiliam, D. (2000). Integrating Summative and Formative Functions Of Assessment. *European Association for Educational Assessment*, (November), 1–25. Prague.

Willse, J. T. (2018). *John T. Willse (2018). CTT: Classical Test Theory Functions*.

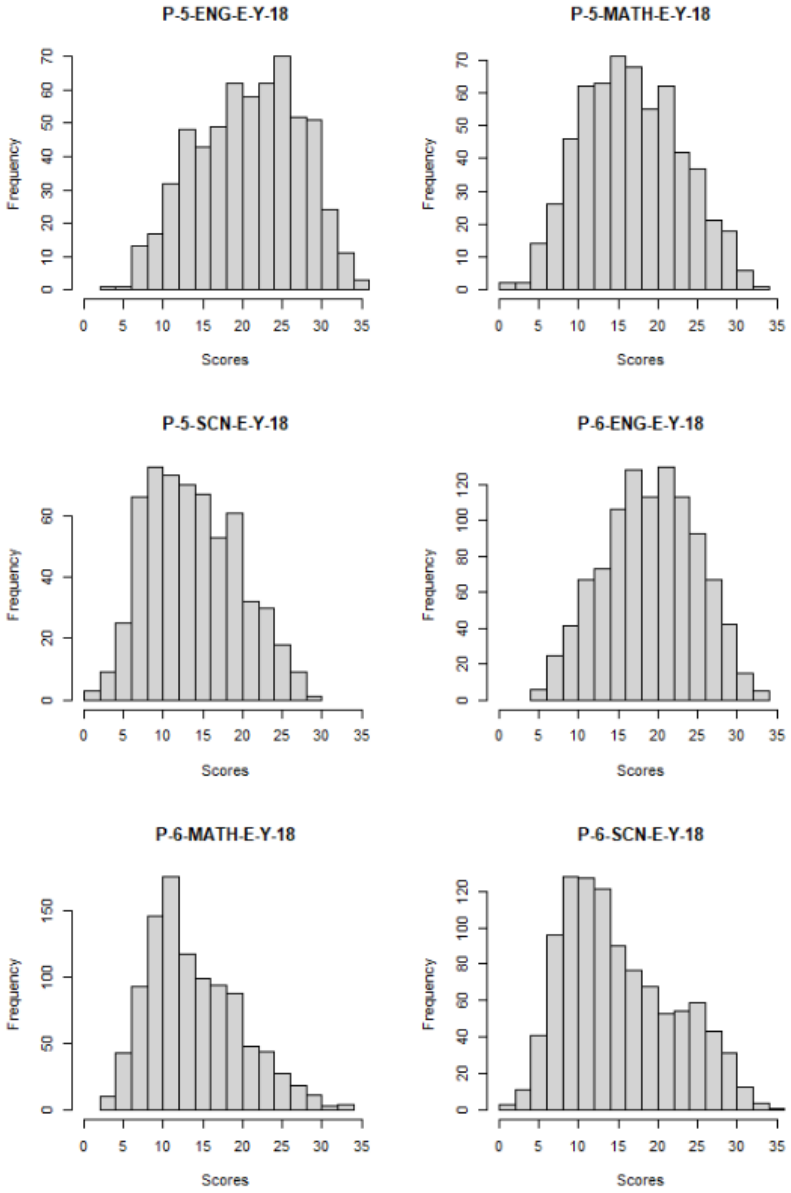
Xiao, Y., Liu, H., & Hau, K. T. (2019). A Comparison of CFA, ESEM, and BSEM in Test Structure Analysis. *Structural Equation Modeling*, *26*(5), 665–677. https://doi.org/10.1080/10705511.2018.1562928

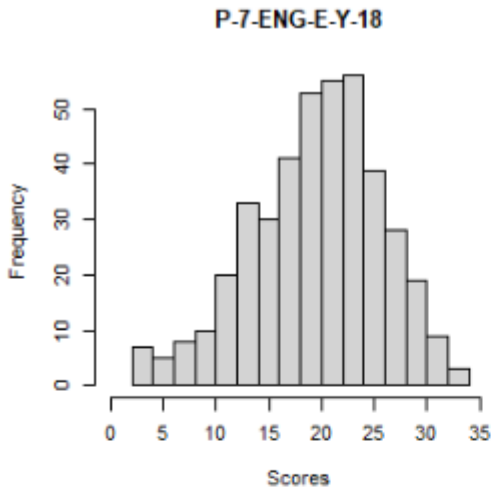
Xu, J., Paek, I., & Xia, Y. (2017). Investigating the Behaviors of M2 and RMSEA2 in Fitting a Unidimensional Model to Multidimensional Data. *Applied Psychological Measurement*, *41*(8), 632–644. https://doi.org/10.1177/0146621617710464

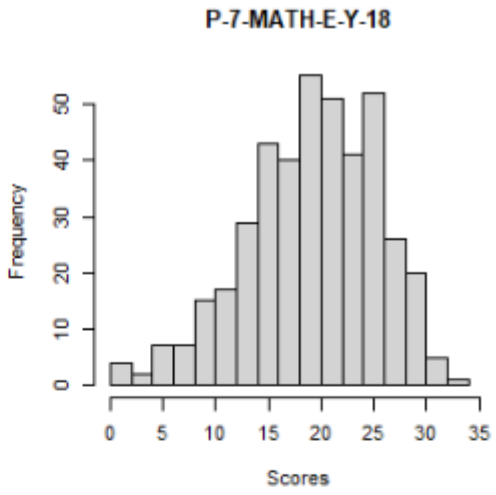
Xu, Y., & Brown, G. T. L. (2016). Teacher assessment literacy in practice: A reconceptualization. *Teaching and Teacher Education*, *58*, 149–162. https://doi.org/10.1016/j.tate.2016.05.010

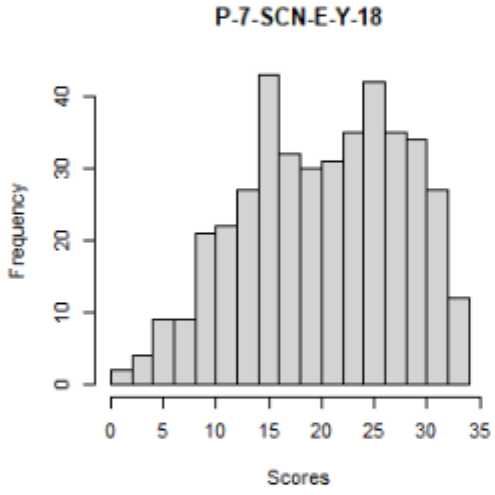
8. Appendices

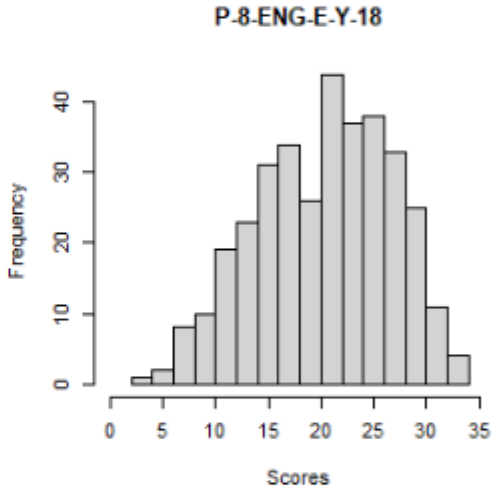
Appendix A Paper Histograms

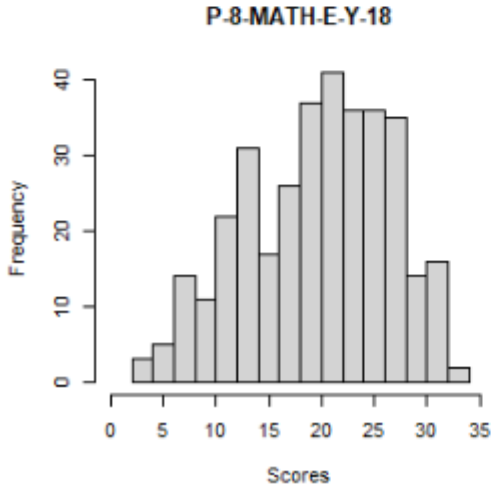


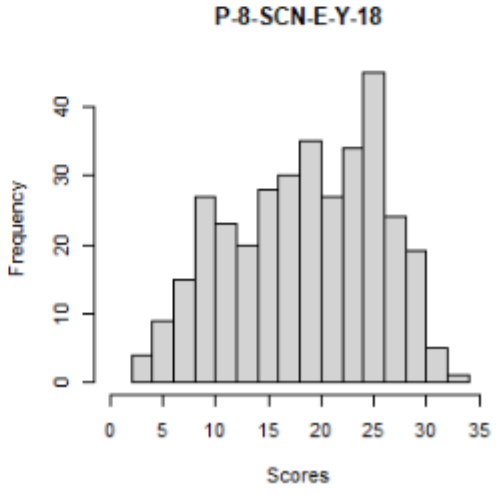

P-7-ENG-E-Y-18
Frequency
0
10
20
30
40
50
0
5
10
15
20
25
30
35
Scores

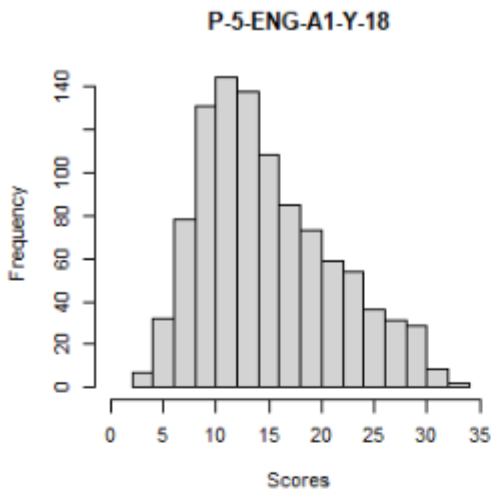

P-7-MATH-E-Y-18
Frequency
0
10
20
30
40
50
0
5
10
15
20
25
30
35
Scores

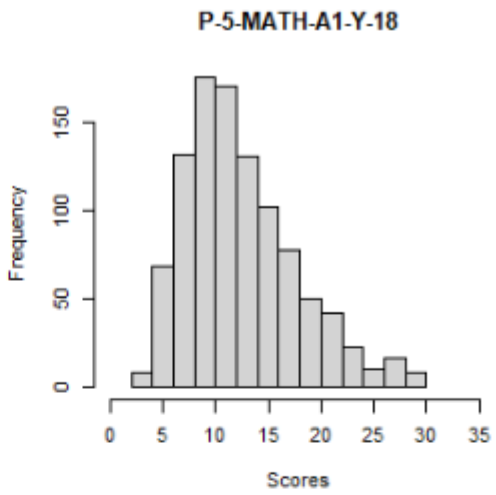

P-7-SCN-E-Y-18
Frequency
0
10
20
30
40
0
5
10
15
20
25
30
35
Scores

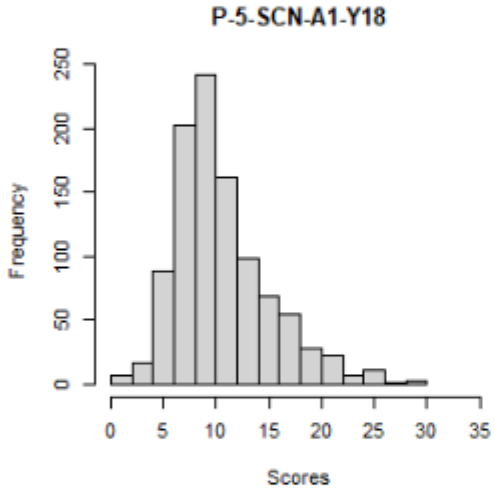

P-8-ENG-E-Y-18
Frequency
0
10
20
30
40
0
5
10
15
20
25
30
35
Scores

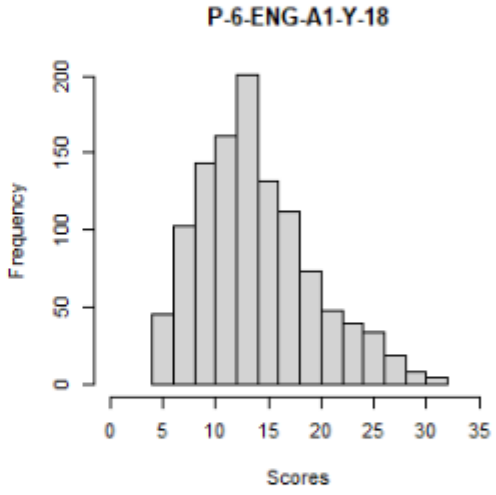

P-8-MATH-E-Y-18
Frequency
0
10
20
30
40
0
5
10
15
20
25
30
35
Scores

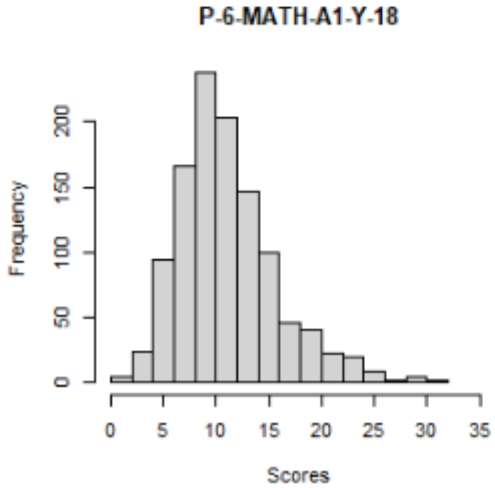

P-8-SCN-E-Y-18
Frequency
0
10
20
30
40
0
5
10
15
20
25
30
35
Scores

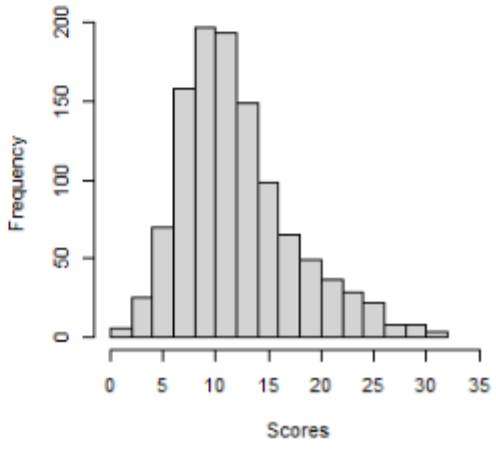

P-5-ENG-A1-Y-18
Frequency
0
20
40
60
80
100
140
0
5
10
15
20
25
30
35
Scores

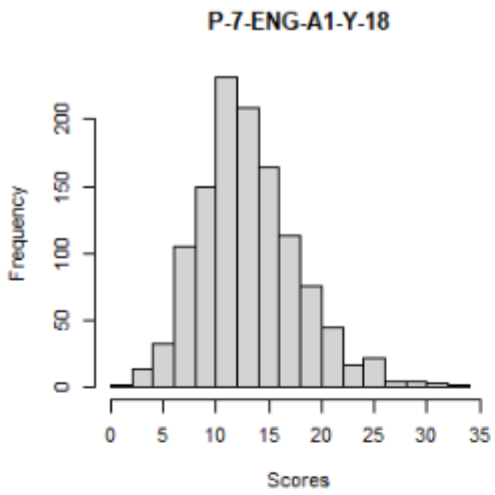

P-5-MATH-A1-Y-18
Frequency
0
50
100
150
0
5
10
15
20
25
30
35
Scores

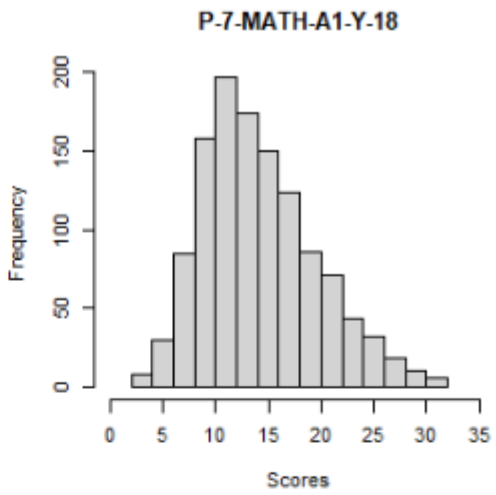

P-5-SCN-A1-Y18
Frequency
0
50
100
150
200
250
0
5
10
15
20
25
30
35
Scores

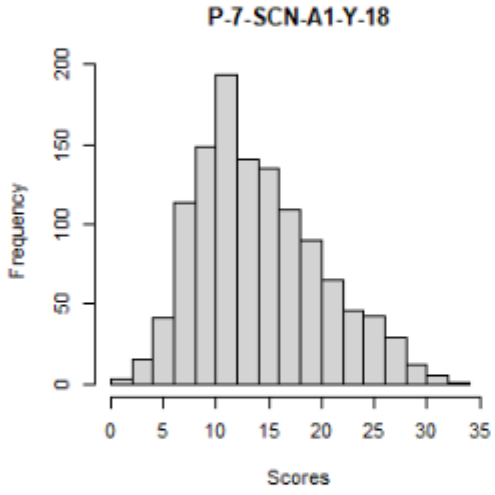

P-6-ENG-A1-Y-18
Frequency
0
50
100
150
200
0
5
10
15
20
25
30
35
Scores

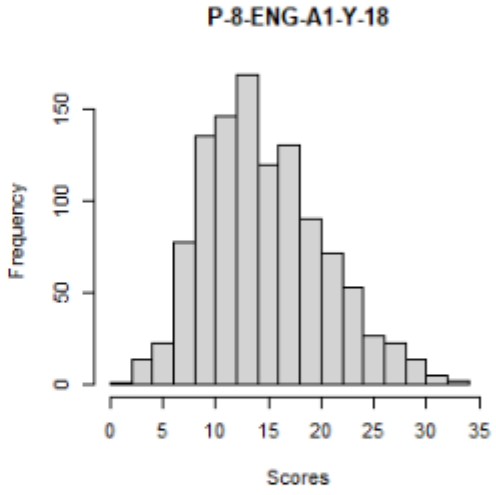

P-6-MATH-A1-Y-18
Frequency
0
50
100
150
200
0
5
10
15
20
25
30
35
Scores

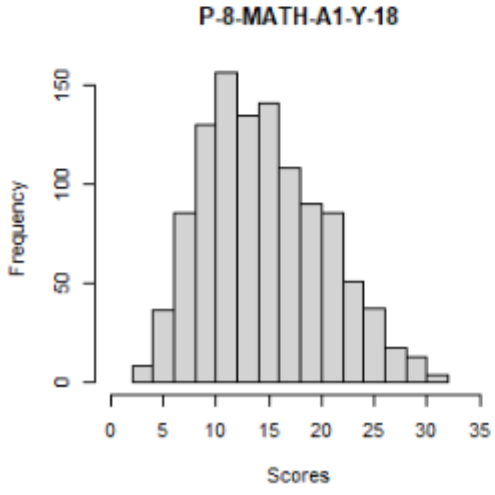

P-6-SCN-A1-Y-18
Frequency
0
50
100
150
200
0
5
10
15
20
25
30
35
Scores

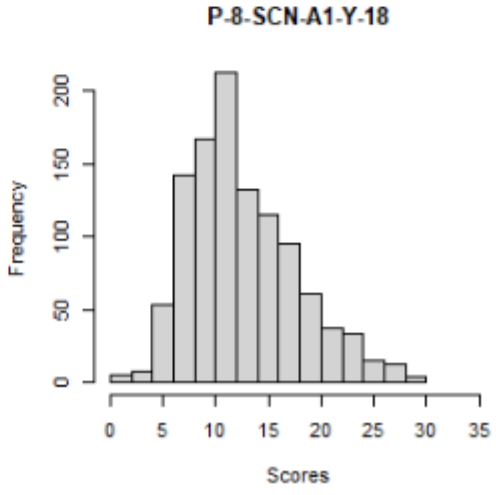

P-7-ENG-A1-Y-18
Frequency
0
50
100
150
200
0
5
10
15
20
25
30
35
Scores


P-7-MATH-A1-Y-18
Frequency
0
50
100
150
200
0
5
10
15
20
25
30
35
Scores


P-7-SCN-A1-Y-18
Frequency
0
50
100
150
200
0
5
10
15
20
25
30
35
Scores


P-8-ENG-A1-Y-18
Frequency
0
50
100
150
0
5
10
15
20
25
30
35
Scores


P-8-MATH-A1-Y-18
Frequency
0
50
100
150
0
5
10
15
20
25
30
35
Scores


P-8-SCN-A1-Y-18
Frequency
0
50
100
150
200
0
5
10
15
20
25
30
35
Scores

Appendix B DISCA Competencies, Skills and Sub-skills

