Cognitive psychometric modelling of students' mathematics performance in the Trends in International Mathematics and Science Study

Jamie Stiff
Department of Education
Kellogg College

Thesis submitted in partial fulfilment of the requirements
for the Degree of Doctor of Philosophy in Education

October 2025

Supervisors:
Professor Jenni Ingram
Dr Michelle Meadows
Dr Stuart Cadwallader

Abstract

Over the past 15 years, educators and educational assessment researchers have shown increasing interest in cognitive diagnostic assessment (CDA) due to its potential to extract more meaningful and pedagogically useful information from test performance. However, analyses employing the cognitive psychometric models (CPMs) used to analyse data from CDAs typically require larger sample sizes than many smaller-scale studies allow, limiting their real-world applications. Several CDA analyses have sought to counteract this limitation by taking advantage of the large sample sizes and ease of access afforded by datasets from international large-scale assessments (ILSAs) such as the Trends in International Mathematics and Science Study (TIMSS).

Much of the existing research using ILSA data has arguably conducted these analyses prematurely. There has been a disproportionate use of CPMs that look to explain students' performance in these studies relative to the use of similar models that seek to explain the psychometric properties of the test-items. Several studies have looked to explain students' performance in TIMSS mathematics, for example, with reference to the attributes of the TIMSS assessment framework. These include the various mathematical topics and cognitive processes that loosely describe the content of TIMSS test-items. However, there has been little-to-no exploratory work in establishing that these attributes explain what individual test-items used in TIMSS measure. If the relationship between attributes and item difficulty is poor or non-existent, it strongly implies that there are other or different driving factors that explain what test-items have assessed. As such, there is a strong rationale to establish how well the TIMSS assessment framework correlates with item difficulty to evaluate the validity of existing CPM studies that have made explanatory claims about students' knowledge and skills in mathematics.

Broader concerns have been raised about the retrofitting of person-diagnostic CPMs to existing assessment datasets. If there is a mismatch between the initial goals of an assessment and the goals of those conducting post-hoc cognitive psychometric modelling on the assessment data, there may be a high likelihood of attribute misspecification. That is, the attributes used by those CPMs to help make explanatory claims about students or test-items run the risk of lacking validity

because those attributes do not relate well to what test-items actually assessed, or because the underlying data does not meet the somewhat strict structural demands of many CPMs.

This integrated thesis is comprised of three papers that explore the viability and validity of cognitive psychometric modelling on data from TIMSS mathematics. The first paper presents a thematic review of CPM research on ILSA datasets. It was found that, relative to the number of articles concerned with modelling students' mastery of different mathematical skills, little had been done to model the item-level attributes that account for difficulty. This raises questions about the validity of these examinee-focused analyses, especially given that much of this analysis has relied on the TIMSS assessment framework as the foundation for their cognitive models.

Having established the need for more item-explanatory modelling, paper 2 makes use of the linear logistic test model to assess how well attributes derived from the TIMSS assessment framework predict item difficulties. Compared to an alternative attribute framework identified in paper 1, the TIMSS framework does a poor job of explaining item difficulties, raising further questions about the suitability of the TIMSS assessment framework for CPM analysis.

However, the third and final paper presents analyses that demonstrate support for the broader concerns regarding CPM retrofitting; neither the TIMSS assessment framework nor the framework that better explained item difficulty in paper 2 successfully explained student performance. The paper presents detailed discussions of several issues associated with the retrofitting process that demonstrate the key issue of attribute misspecification and underlying structural issues with the TIMSS dataset. It is argued in the paper that these are fundamental issues relating to the breadth of test-item content in TIMSS mathematics relative to the number of items completed by students. This renders it impossible to construct a sufficiently accurate account of what test-items have measured whilst simultaneously meeting the demands of the CPMs.

This thesis demonstrates that item-level data from TIMSS are not suitable for cognitive psychometric modelling, particularly that aimed at explaining student performance, and highlights broader risks of retrofitting CPMs to ILSA datasets without first establishing item-attribute validity.

Acknowledgements

I would like to thank the many people who have made the completion of this thesis possible.

First, I must thank my wonderful supervisors Professor Jenni Ingram, Dr Michelle Meadows, and Dr Stuart Cadwallader. All three joined me part-way through my DPhil journey and brought a wealth of warmth, humour and experience at a period where I could not envisage any path to completion. That I am here now, submitting a thesis I am proud to put my name to is a testament to their expertise and generosity. I would also like to thank all my other OUCEA colleagues and friends, past and present for their collaboration and support.

I must also give thanks to the friends and family who kept me sane, kept me smiling, and kept me on track. Thank you to my incredible Somervillians, and all my other friends I've made in my 14(!) years here in Oxford. Thank you to my wonderful parents Carol and Chris, and my fantastic big sister Carly. I promise to come home more often!

And lastly, I must thank Mike - perhaps the only person more relieved than I am that this thesis is finally complete! Thank you so much for everything. I couldn't have done this without you.

Contents

1 Introduction

1.1 Cognitive diagnostic assessment

Almost all school-based tests share an overarching goal of providing information about what a student or group of students know or can do in a given subject area. While some assessments may lead to grading, many aim to be in some way diagnostic, helping to identify areas where students may need further support.

However, this link between test performance and further learning is not always made, or, if made, is not always well supported by data from the test itself; it is often not easy to draw strong, specific inferences about students' knowledge and skills solely from their test performance. This is, in part, due to difficulties in making direct links between the skills and knowledge that test-items *intend* to assess, and the skills and competencies that they *actually* measure.

Cognitive diagnostic assessment (CDA) is an approach to assessment that attempts to make these kinds of strong, detailed inferences about students' knowledge and skills. These inferences are drawn from students' patterns of correct and incorrect responses to the collective body of test-items that form an assessment. Each test-item used in a CDA is designed specifically to measure some combination of skills or knowledge states from a wider set of related skills in a topic area. A student's pattern of correct and incorrect responses to the items in a test is then used to make inferences about which of these skills have and have not been mastered (Leighton & Gierl, 2007; Williamson, 2023; Rupp, 2007).

To provide accurate and meaningful interpretations of examinees' skills and knowledge through cognitive diagnostic assessment, there must be a very clear and well-defined relationship between the model of the skills and knowledge that are intended to be measured by a test, commonly referred to as 'attributes', and how these are reflected and measured by each individual item comprising that test. Cognitive diagnostic models (CDMs) are forms of multidimensional latent variable models that, as the name suggests, attempt to 'diagnose' examinees into profiles of mastery and non-mastery of the skills and knowledge under evaluation.

Although CDMs have been seen as attractive options for drawing more pedagogically diagnostic information from assessments, there have been a number of barriers, theoretical and practical, that have limited their implementation. Ravand and Baghaei (2020) commented on several of these barriers, noting that CDM analyses are typically inaccessible to the broader audience that would be interested in conducting them, findings from the better fitting CDMs can often border on uninterpretable in real-world application, and, perhaps most notably, that CDM analyses often require greater sample sizes than many of the practical applications allow.

The development of the field of cognitive diagnostic assessment has primarily been driven by simulation studies, and, albeit to a lesser extent, ‘retrofitting’ of cognitive diagnostic models to existing assessment datasets (Ravand & Baghaei, 2020). While Ravand and Baghaei refer to retrofitting of CDMs to existing assessment datasets as a “measure-of-last-resort purpose” for using CDMs (p.27), they noted that retrofitting studies outnumber ‘real’ cognitive diagnostic assessment studies by around four-to-one.

The ideal process for CDA would be to develop a cognitive model of specific knowledge and skills intended to be assessed, provide a very clear mapping of these onto the development of test-items (an item framework), ensuring sufficient coverage and interactions between attributes of the cognitive model, and then conduct confirmatory analyses. With retrofitting, the cognitive model and item framework often has to be generated post-hoc and applied to test-items that were designed with alternative intentions.

Gierl and Cui (2008) identified several issues associated with retrofitting. This includes the issue of retrofitting multidimensional models to tests that have been designed around unidimensional constructs, tests covering distinct but strongly-related content areas, or tests which otherwise centre on the estimation of a singular examinee ability score. If an assessment has been designed to measure a very limited range of constructs relative to the intended list of knowledge and skills comprising the cognitive model of the retrofitting analysis, then data from that assessment is unlikely to be suitable for retrofitting. Additionally, they argue that retrofitting will typically yield poor model fit to the data, in turn producing inaccurate diagnostic inferences.

Findings from CDM analyses that have been retrofit to existing assessment datasets may therefore lack validity if they are based on item frameworks that relate poorly to the test-items themselves.

One method for assessing the validity of an item framework, either in 'real' CDA or in retrofitting analyses, is to explore how well that framework predicts item difficulty. The attributes that define differences between items in a test should reasonably account for the differences and ranges of their difficulty (Harrison et al., 2023). If an item framework is used to define the process for the development of items, as well as how student performance is scored and interpreted, then there should be a strong relationship between attributes and difficulty. If the relationship between attributes and item difficulty is weak or non-existent, other factors must account for differences in the difficulties of test-items. These may be other, construct-relevant factors that have been unaccounted for in the framework, or construct-irrelevant factors that ideally would be eliminated or minimised during test-item development. In either case, these factors *should* be accounted for, either in the framework, or as a caveat in the interpretation of results. Regardless, identifying the sources of difficulty in test-items is important for establishing construct validity of educational assessments (Embretson, 1984; Messick, 1993), as well as assessing the validity of item frameworks used in retrofitting analyses.

1.2 Cognitive psychometric models

This thesis adopts Rupp's (2007) use of the term 'cognitive psychometric modelling' (CPM) to refer to the collective body of explanatory statistical models concerned with the analysis of item-level data from educational assessments and the interactions between items and examinees. For the purposes of this thesis, it can be seen as an umbrella term for two related groups of models: those that fall under item-response theory (IRT) models, and those that would be categorised as cognitive diagnostic models (CDMs) commonly used by those interested in CDA. While many CDMs are primarily focused on exploring what item-level performance tells us about examinees, there are many IRT-rooted models that place greater emphasis on modelling what examinees' responses tell us about test-items.

Some previous research has broadened the term cognitive diagnostic modelling (CDM) to be inclusive of IRT approaches. That is, traditional IRT models such as the 1PL/2PL/3PL models are treated as a subset of cognitive diagnostic models. While it is helpful to consider item-focused and examinee-focused modelling as complementary rather than distinct, and therefore warrant shared terminology, this thesis follows Rupp's argumentation that the 'diagnostic' part of the CDM term is limiting because much of the psychometric and/or IRT-rooted literature has little concern with any form of diagnosis, at least in isolation.

A common feature shared by the vast majority of CPMs is the use of a Q-matrix. A Q-matrix is an item-by-attribute matrix which defines the relationship between items and the attributes that they intend to assess, or, in the case of retrofitting analyses, are judged to have assessed.

Suppose that a CPM was used to model performance on a fractions test. One attribute of the Q-matrix might, for example, define whether each test-item included an improper fraction, while another might define whether each question tested a division of an integer by a fraction. Other attributes might cover properties of how the item was administered to examinees, such as whether the question was multiple-choice or whether the question included a diagram. The Q-matrix is therefore one of the fundamental components of these CPMs that shapes what sorts of psychometric or diagnostic parameters are drawn from the item-level performance data, and the kinds of conclusions that can be made.

An example of how a Q-matrix might be constructed is shown in Table 1. In most CPM analyses, Q-matrices represent either presence (1) or absence (0) of each an attribute, though variants that weight each attribute for each item can also be used.

Table 1: Example of a Q-matrix

Item	Attribute j_1	Attribute j_2	Attribute j_3	Attribute j_4	Attribute j_5	Attribute j_6
i_1	1	0	0	0	0	0
i_2	0	1	0	0	0	0
i_3	1	1	0	0	0	0
i_4	0	0	1	0	0	0
i_5	0	1	1	0	0	0
i_6	0	0	0	1	0	0
i_7	1	0	0	1	0	0
i_8	0	0	0	0	1	0
i_9	0	0	0	1	1	0
i_{10}	0	0	0	0	0	1
i_{11}	1	0	0	0	0	1
i_{12}	0	0	1	0	1	0
i_{13}	0	1	0	1	0	0
i_{14}	1	0	1	0	0	1
i_{15}	0	0	1	0	1	1

Misspecification, either because the attributes used in the item framework to construct the Q-matrix are poorly related to the items generally or because the attributes were incorrectly assigned to items within the Q-matrix, can have large knock-on consequences to the results of these analyses; Rupp (2007) stated that Q-matrices in CPM analyses that have not been comprehensively justified can produce a “garbage-in, garbage-out” phenomenon (p. 98). Consequently, the selection of attributes to be used in a Q-matrix, and the correct assignment of attributes to items is one of the most important, and most challenging processes of a CPM analysis.

1.3 The Trends in International Mathematics and Science Study (TIMSS)

One of the most influential of the international large-scale assessments (ILSAs) is the Trends in International Mathematics and Science Study (TIMSS). TIMSS, alongside its sister-study, the Progress in International Reading Literacy Study (PIRLS), are directed by the International Association for the Evaluation of Educational Achievement (IEA). TIMSS has been conducted

every four years since 1995 and assesses and compares the mathematics and science performance of students in different countries around the world, targeting two different age groups – students in the fourth international grade of education (ages 9-10) and students in the eighth grade (ages 13-14). In each participating country, large, nationally representative samples of students are drawn to participate in the subject assessments, and supplementary information is collected through the use of questionnaires directed at schools and parents, as well as the students themselves. While earlier cycles of the study were conducted with pencil and paper, as of the 2023 cycle, TIMSS is administered digitally.

Shortly after the release of TIMSS results, the datasets are made publicly available for researchers and other interested parties to explore further. The ease of access to large, nationally representative samples of item-level performance in the assessments, supported by a wealth of additional contextual information from the questionnaires makes ILSA data an attractive option for a wide array of research questions. This includes those interested in retrofitting CPMs to ILSA datasets. Relative to other ILSAs and subject areas, TIMSS mathematics has drawn substantial attention from researchers applying CPMs (see paper 1, Chapter 4). This may be because the test-items used in TIMSS mathematics, relative to the mathematics test-items used in the Programme for International Student Assessment (PISA), are based on a relatively simple framework of mathematical content and cognitive processes that can be easily mapped onto individual test-items. This makes TIMSS appear, at a surface level, well-suited to cognitive psychometric modelling.

Much of the existing research utilising data from TIMSS mathematics for CPM purposes has made use of the TIMSS assessment framework, which defines the relationship between items and facets of mathematics that TIMSS aims to assess (see paper 1, Chapter 4). What has not been well established is how well this framework actually relates to examinees' interactions with these items, and in particular, how attributes of the item framework relate to the difficulty of test-items. It appears that the relationship between the TIMSS assessment framework and TIMSS item difficulty has been assumed without sufficient analyses of whether the framework fits the data, and therefore, whether these retrofitting analyses are appropriate and valid.

1.4 Assessment frameworks and item difficulty

In a well-targeted test, items should vary in difficulty to reflect the range of students' abilities. The differences in the difficulties of these items should, ideally, be driven by intended and understood subject-relevant sources of item difficulty (Pollitt et al., 2007; Messick, 1993).

While TIMSS mathematics assesses and scores students at an overall level of mathematics, as defined by the TIMSS assessment framework, there is an understanding that different content areas and cognitive skills collectively contribute to what subject experts would consider 'mathematics', and that different students or groups of students might have varying strengths and weaknesses in those areas. Although the IRT approaches used for scaling performance and estimating item parameters, such as item difficulty, are based on unidimensional models, there is an assumption that these scaling approaches are valid for content subscale reporting because the content and cognitive domains collectively form an overarching unidimensional construct (von Davier, 2020). The TIMSS assessment framework provides greater detail as to more specific topics within the content areas, as discussed briefly in section 1.3, though no further scaled scoring is conducted by the IEA at this level of specificity. The IEA do not claim that the TIMSS assessment framework explains all of the sources of difficulty of their test-items, and have previously recognised that their mathematics test-items could include unintended sources of difficulty, such as those arising from reading demands (Martin & Mullis, 2013).

The question may be raised as to whether the validity of the results of TIMSS, as reported by the IEA, are threatened by a weak relationship between the specific mathematical attributes discussed in the assessment framework and the factors driving the difficulties of TIMSS test-items.

TIMSS reports on mathematics (and science) performance broadly in terms of overall scores, and scores in major content domains (e.g. number, geometry, data; von Davier et al., 2024). If the difficulty of the test-items included in TIMSS are not driven by mathematically relevant factors (i.e. construct-irrelevant factors), or there is evidence that performance on TIMSS mathematics items is not well explained by a unidimensional construct, this would pose a threat to the validity of

TIMSS overall score reporting. If the difficulty of the test-items included in TIMSS *are* driven by mathematically-relevant attributes of those items, and there *is* sufficient evidence of unidimensionality, then the conclusions reported by the IEA about mathematics performance generally are relatively well supported by the item-level data. In other words, the kinds of conclusions drawn in the IEA's reporting of TIMSS results may be relatively robust against minor attribute misspecification.

However, as the level of specificity in the claims made about students' knowledge and skills increases, so too does the need for there to be a strong link between the claims made about what those test-items have measured, and what they have *actually* measured (Kane, 2016; Messick, 1993). This is of particular relevance when evaluating the validity of cognitive diagnostic modelling studies that make highly specific inferences about students' cognitive states from item-level data.

For example, suppose a mathematics test consisted solely of multiplication questions and a goal of the assessment was to identify students that would benefit from specific support with multiplications involving decimal numbers. If subsequent analysis reveals that, despite the test's overall appropriate difficulty level, there is no correlation between the presence of decimal numbers in individual items and their difficulty levels, then it becomes apparent that other factors are influencing item difficulties. While this does not necessarily invalidate the test as a measure of general multiplication skills, it does undermine the validity of any conclusions drawn about students' specific struggles with decimal-based multiplication.

Given that previous cognitive diagnostic modelling studies have lifted attributes from the TIMSS assessment framework to make these kinds of specific comments about students' strengths and weaknesses in mathematics (e.g. Lee et al., 2011; Yamaguchi & Okada, 2018; Delafontaine et al., 2022), there is a clear need to first establish whether the framework even relates to item difficulty. Further, there is a need to establish whether the datasets from international assessments are more generally suitable for these kinds of high-specificity diagnostic inferences.

1.5 Research questions

This thesis will look to answer the main overarching question:

Can cognitive psychometric modelling approaches be viably and validly applied to TIMSS data to provide diagnostic information about students' knowledge and skills in mathematics?

To assist in answering this overarching question, three journal-style research papers are presented, herein referred to as paper 1, paper 2, and paper 3. Each one of these papers addresses its own research sub-question.

- ***Research sub-question 1 (paper 1):***

  What are the main themes, foci, and findings of research utilising international assessment datasets for cognitive psychometric modelling purposes?

- ***Research sub-question 2 (paper 2):***

  Does the TIMSS mathematics framework possess sufficient predictive power to explain the item-level factors contributing to item difficulty, and if not, how could the framework be improved upon?

  Research sub-question 3 (paper 3):

  Are the item-level data available from TIMSS suitable for cognitive diagnostic modelling of student performance, and if not, what are the threats to the validity of these analyses?

These three research sub-questions collectively address the overarching thesis research question by exploring how CPMs have been used to analyse ILSA data in other studies, establishing what accounts for the difficulties of test-items in TIMSS mathematics, and identifying what issues arise when applying CDMs to TIMSS data.

1.6 Outline of the integrated thesis

Chapter 1 has presented a brief overview of the context of the thesis, including an introduction to cognitive diagnostic assessment, cognitive psychometric models, and the TIMSS assessment

context. It has also introduced the main overarching question guiding the thesis, as well as three research sub-questions that will be the focus of each of the three papers.

Chapter 2 presents detailed discussions of three conceptual underpinnings that are central to the thesis. The chapter begins by presenting the content and structure of the fourth-grade mathematics assessment used in TIMSS (with particular focus on TIMSS 2011, the cycle of the study that data are taken from in paper 2 and paper 3). The chapter then discusses what is meant by the term 'cognitive psychometric models' and why this term was selected for the thesis, before presenting some key examples of cognitive psychometric models and the debates surrounding their use. The chapter concludes with a discussion of item difficulty in educational assessments and the challenges associated with predicting item difficulties from their content.

Chapter 3 provides an overview of the methodological approach taken in each of the three papers. Each paper targets one of the three sub-questions, and this chapter explains how these, together with the overarching research question, are addressed.

Chapter 4 focuses on the first of the three papers of the integrated thesis, and primarily targets **research sub-question 1** (***What are the main themes, foci, and findings of research utilising international assessment datasets for cognitive psychometric modelling purposes?***). This question is addressed through the presentation of a thematic review of research using CPMs to explore datasets from the three major ILSAs of school-age students – TIMSS, PISA, and PIRLS. This paper establishes common themes in the existing body of literature, such as which studies and subjects have garnered the most interest from researchers interested in cognitive psychometric modelling, what are the main CPMs that have been employed, and what attribute frameworks have been used to parameterise performance. The results are synthesised into a qualitative discussion of the state of CPM applications to ILSA data, and a key weakness in this body of literature is identified: the scarcity of item-explanatory modelling of TIMSS mathematics relative to the wide range of person-diagnostic analyses. This establishes the rationale for papers 2 and 3 which aim to address this gap in the literature.

Chapter 5 presents the second of the three papers and primarily targets **research sub-question 2** (***Does the TIMSS mathematics framework possess sufficient predictive power to explain the item-level factors contributing to item difficulty, and if not, how could the framework be improved upon?).*** Having established in paper 1 that there is a need for more research that explores item difficulty in ILSAs such as TIMSS, paper 2 analyses how well the TIMSS mathematics assessment framework predicts the difficulty of its test-items. The analysis focuses on the performance of approximately 20,000 fourth-grade students from six Anglophone countries that participated in TIMSS 2011 and looks at their responses to a collection of 85 items, around half of the fourth-grade mathematics assessments included in the study. A combination of the one-parameter logistic model and the linear logistic test model (LLTM; Fischer, 1973) are used to calculate estimates of item difficulty. These specific CPMs are discussed in greater depth in both Chapter 3 as well as the methods section of the paper. The paper explores the relative fit of the data for these six countries to the international dataset and compares the correlations and relative fit of item difficulty estimates produced under the two models. Model fit of the LLTM under four different Q-matrices is explored and compared to that of the 1PL. Three of these Q-matrices are adapted from the TIMSS 2011 assessment framework, while the fourth is adapted from the Q-matrix originally used by Tatsuoka et al. (2004), one of the most widely cited person-diagnostic analyses of TIMSS mathematics that was identified and discussed in paper 1. All three of the Q-matrices based on the TIMSS assessment framework show poor fit to the 1PL item difficulty estimates, with the 1PL showing significantly greater fit to the data, indicating that the Q-matrices based on the framework poorly explain the item-level attributes that contribute to their difficulty. The alternative Q-matrix correlates more strongly and has a better fit to the underlying performance data though still fails to capture the majority of variance in item difficulties.

Chapter 6 presents the third and final paper of the integrated thesis, and primarily targets **research sub-question 3** (***Are the item-level data available from TIMSS suitable for cognitive diagnostic modelling of student performance, and if not, what are the threats to the validity of these analyses?).*** The paper investigates the suitability of TIMSS data for person-diagnostic modelling by presenting an example application of the DINA model to item-

level performance data from TIMSS 2011. Two Q-matrices are used in the analysis: one based on the TIMSS content and cognitive domains, and the second based on the alternative Q-matrix that performed best in the item-explanatory analysis of paper 2. The analysis presented in the paper reveals that neither the TIMSS framework Q-matrix nor the alternative Q-matrix supports valid person-diagnostic inferences, with widespread item misfit and implausible mastery classifications. Issues relating to valid Q-matrix construction are discussed. It is argued that the breadth of test-content relative to the number of completed items at the student-level, in conjunction with the demands of the DINA model, makes it impossible to avoid Q-matrix misspecification. The paper raises concerns more broadly about the suitability of ILSA datasets for person-diagnostic analyses.

Chapter 7 synthesises findings from across the three papers. Collectively, the three papers show that despite the common use of the TIMSS assessment framework as the basis for explaining student performance, the attributes of the assessment framework do not explain the functioning of items well. This questions the validity of previous analyses that have relied on the assumption that these attributes explain performance on TIMSS test-items. However, even if alternative attributes were selected, it is unlikely that *any* framework of suitable attributes could be applied to TIMSS test-items; the breadth of test-content is too wide relative to the number of test-items completed by individual students. Chapter 7 reflects on the overarching thesis question and concludes that **there is very limited potential for cognitive psychometric modelling approaches to draw valid explanatory conclusions about students' knowledge and skills in mathematics.** This is because there is a fundamental mismatch between the goals of person-diagnostic analyses and the assessment design of TIMSS. Some potential limitations of the thesis' approach are discussed, and where appropriate, counterarguments to these are presented. The chapter also discusses a potential idea for further item-explanatory research utilising TIMSS assessment materials in independent samples.

Supplementary information on the analyses presented in papers 2 and 3 are available within the appendices presented after the thesis reference list and list of thesis tables and figures. The brief summary of these appendices is as follows:

- **Appendix A** – a copy of the approval to access and use the IEA's restricted-use items in TIMSS 2015
- **Appendix B** – a copy of the instructions given to reliability coders for Q-matrix D (paper 2)
- **Appendix C** – all six complete Q-matrices used in this thesis (Q-matrices A-D and Q-matrices 1-2).
- **Appendix D** – Item-fit statistics (paper 2)
- **Appendix E** – A comparison of model fit between Q-matrix D (paper 2) and Q-matrix 2 (paper 3) under the linear logistic test model, to explore potential loss of model fit under attribute reduction.
- **Appendix F** – Item-level guessing and slipping parameters (paper 3)

2 Conceptual underpinnings

This chapter focuses on providing discussion on key concepts, models, and previous research findings relating to three of the overarching topics central to this thesis;

- **The TIMSS mathematics assessment**, and in particular, the assessment framework and item design used in TIMSS.
- **Cognitive psychometric modelling**, including a discussion on the selection of this terminology and the item-explanatory models and person-diagnostic models that fall under this umbrella term.
- **Item difficulty**, including how it has been conceptualised, what item-level factors have been associated with item difficulty both in mathematics and more generally, and a discussion on why predicting item difficulties reliably remains an ongoing challenge.

2.1 TIMSS

2.1.1 The origins and purpose of TIMSS

TIMSS is the world's longest running international educational assessment (von Davier et al., 2024), first being carried out in 1995 and then every four years thereafter. While the assessment structure has changed slightly with each iteration of the study, TIMSS has maintained its original goal of providing internationally comparative data about trends in students' performance in mathematics and science. In the IEA's international reporting of the results of TIMSS 2023 (von Davier et al., 2024; https://timss2023.org/results/), they stated that TIMSS results "allow countries to benchmark their performance against other nations, fostering a global dialogue on best practices in mathematics and science education" and that TIMSS "facilitates evidence-based decisions about education policy and resource allocation". The results of each study provide a broad overview of performance standards in mathematics and science in participating countries. TIMSS also reports on students' attitudes, beliefs and behaviours towards mathematics, science and school generally, collected through supplementary questionnaires.

Since the third cycle of the study in 2003, TIMSS has assessed students in both the fourth and eighth international grades[1]. As the ages at which students start schooling vary between

[1] TIMSS 1995 also assessed students in both grades. The 1999 study only assessed students in the eighth grade.

countries, this sometimes corresponds to different national grades, for example, years 5 and 9 for participating students in England. In most cases, participating students will be between 9-10 years old and 13-14 years old respectively. Samples are drawn from all participating countries using stratified sampling methods designed to ensure that the cohorts of students selected for TIMSS in each country are nationally representative, both in terms of academic performance and broader socioeconomic characteristics.

All students are assessed in both mathematics and science, completing one booklet of test-items for each subject, typically in two separate sessions in the same day. Students then complete their questionnaire later that day. The analyses presented in paper 2 and paper 3 of this thesis focus specifically on the mathematics assessment completed by students in the fourth international grade, though references to the eighth-grade assessment, science assessments, and questionnaire data will be made as appropriate.

The IEA report that TIMSS is designed specifically to "provide a comprehensive picture of the mathematics and science achievement of fourth- and eighth-grade students in each participating country" (Mullis et al., 2021; p.73). Notably, TIMSS prioritises the accurate estimation of performance in mathematics and science at the *national* level, rather than at the *individual* level. This goal is reflected in the ways in which students are assessed, and in how performance is scored.

Each fourth-grade student that participates in TIMSS completes a booklet of mathematics test-items. This booklet consists of approximately 25 test-items. However, TIMSS administers far more test-items than are completed by any one student; the fourth-grade mathematics assessment used in TIMSS 2011, for example, consisted of a total of 177 administered items. This is, in part, because TIMSS looks to assess a wider range of mathematics content than would be reasonable to assess in a single session. To minimise the demands placed on individual students, TIMSS administers items using a multi-matrix sampling approach (Mislevy et al., 1992; Mullis et al., 2019) in which test-items are distributed across blocks, and blocks across booklets. This design allows for the different booklets of items to contain different (but systematically

overlapping) blocks of items while allowing all item-level and person-level parameters to be calculated on a single scale. This facilitates comparisons between groups of students who have completed entirely different test-items to one another.

For each of mathematics and science, TIMSS reports country-level performance on an overall competency scale, with a mean score of 500 and standard deviation of 100. This scaling is based on the average mathematics and science performance of the countries that participated in TIMSS 1995, and scores of more recent cycles have been benchmarked on this same scale to facilitate longitudinal comparisons of national performance.

A set of five 'plausible values' (Mislevy, 1991) are generated for each student, which provide estimates as to where each student's 'real' ability may lie on the competency scale. These plausible values are randomly drawn from a conditional normal distribution based on each examinee's item-level response data, but also from contextual information derived from their questionnaire responses[2]. This means that two students with identical responses to the same set of test-items will have different plausible values, both because the draws are random, and also because their questionnaire responses affect those draws. Previous research has suggested that the inclusion of this contextual information in plausible value draws helps to reduce bias in group-level comparisons (e.g. Mislevy, 1991; von Davier et al., 2009). In the TIMSS 2023 technical report (Bezirhan & von Davier, 2024), they reiterated that these plausible values are not designed to be used for individual student reporting and should only be used for estimating the performances of large groups.

While the analyses presented in the papers of this thesis do not rely on these plausible values, their use in TIMSS warrants discussion because this methodology is consistent with the IEA's goals for the study. TIMSS is designed for the estimation of the performance of populations (i.e. countries, or large sub-groups within those countries) in two broad subject domains. This contrasts starkly with the goals of person-diagnostic modelling, which are more concerned with

[2] Questionnaire data used in plausible value estimation is contextualised at the national level, rather than at the international level.

the performance of individuals, and typically within much more granular topic areas (Rupp, 2007; Williamson, 2023).

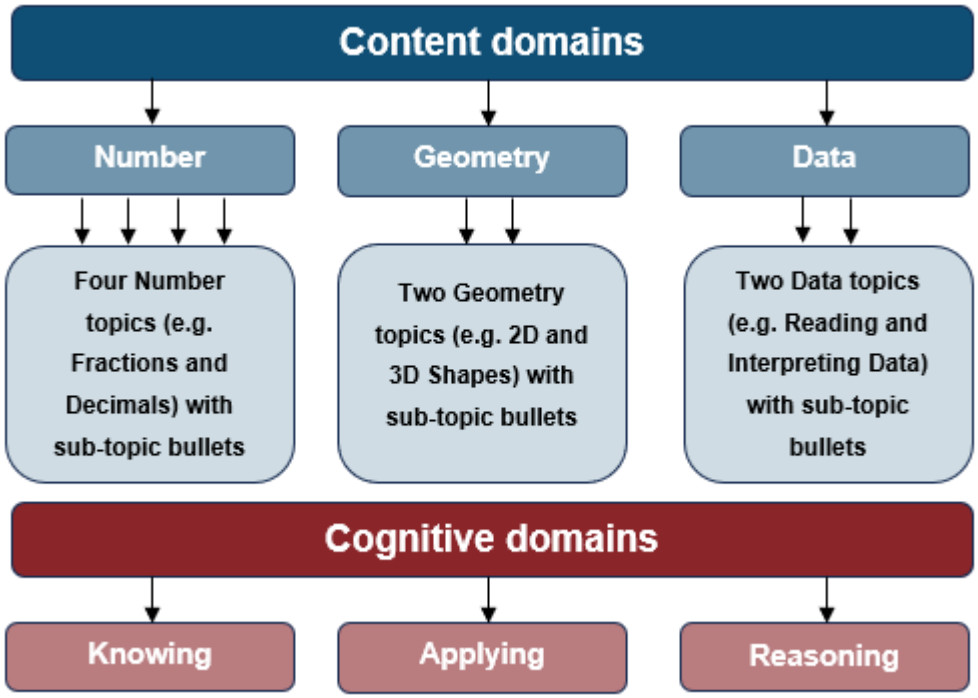
Advocates of CPMs might argue that these explanatory models could extract more pedagogically useful information from ILSAs like TIMSS. However, as discussed, the assessment designs of these studies are specifically designed for drawing population-level inferences, and may even actively discourage student-level inferences.

2.1.2 TIMSS assessment frameworks and structure

While TIMSS places greater emphasis on overall score reporting, the IEA do also report breakdowns of performance within the major content areas and cognitive processes associated with groups of test-items, as defined within the TIMSS assessment framework (e.g. Mullis et al., 2021). These are used to provide insights into where the differences in overall scores between groups arise, or provide information about differences in performance between groups that are otherwise masked at the overall performance level.

The content and cognitive domains used in the 2011 cycle of TIMSS mathematics (Mullis et al., 2009) are reported in Figure 1. The division of content areas and cognitive processes is one of the defining characteristics of the TIMSS assessment frameworks. These content and cognitive domains outline both how the test-items are constructed (in terms of what students are asked to do, and how), and how performance at the national level is reported. The assessment frameworks used in TIMSS have changed slightly over cycles, and from the 2019 cycle onwards, TIMSS has transitioned into a digitally administered assessment, with no paper-based assessment in 2023. As two of the three research papers presented in this thesis use item-level performance data from the 2011 cycle, that cycle is therefore the focus of the discussion here, though many of the main discussion points still apply as of the 2023 cycle.

Figure 1: Summary of content and cognitive domains in TIMSS 2011



In the TIMSS 2011 fourth-grade mathematics assessment, each test-item was categorised as assessing one of the three major content domains of the fourth-grade assessment framework – 'number', 'geometric shapes and measures' (herein, 'geometry'), and 'data display' (herein, 'data'). Further subdivisions were made into the topics included in each of the content domains. For example, there were four topics within the number content domain in TIMSS 2011; whole numbers, fractions and decimals, number sentences, and patterns and relationships. Each of these also had an associated list of 'topic bullets' outlining more specific content assessed within each topic. For instance, the 'fractions and decimals' topic had six associated bullets, such as "add and subtract simple fractions" and "show understanding of decimal place value, including representing decimals using words, numbers, or models". Every item was categorised as assessing a single topic bullet, belonging to a single topic that belonged to a single content domain, though the IEA only reported on and provide subscale scoring at the content domain level.

Each item in TIMSS 2011 also assessed one of three cognitive domains. These represent the types of cognitive skills that students needed to draw on to reach the correct answer. These three cognitive domains were 'knowing', which relates to the ability to recall mathematical definitions

and simple number and geometry properties, and to perform routine calculations, 'applying', which requires students to make simple choices about mathematical operations or to solve routine problems, and 'reasoning', in which students have to make connections between different elements of mathematical knowledge, make justifications and generalisations, and solve non-routine problems. While the TIMSS 2011 mathematics framework listed the kinds of processes associated with each cognitive domain (e.g. 'recall', 'recognise', 'compute' for the 'knowing' domain), the sets of these measured by individual items was not reported at this level of granularity, as was the case with the topic bullets in the content domain.

Differences in overall performance between countries or students can be contextualised by assessing performance on the different content and cognitive domain subscales. For example, in TIMSS 2019, the significant increase in England's overall TIMSS mathematics score from previous cycles had been driven by significant improvements in students' performance in the number and data domains, whereas there had been no change in scores in the geometry domain (Richardson et al. 2020). Performance in geometry was significantly lower than in the other content domains. Similarly, while performance in 2019 had significantly improved from 2007 and 2011 across all three cognitive domains, the improvements were greatest in the reasoning domain, which had previously been an area of relative weakness for students in England, and was the only cognitive domain where England's score had significantly improved from 2015.

2.2 An introduction to cognitive psychometric models

2.2.1 Why cognitive 'psychometric' models?

As discussed in section 1.2, this thesis uses the term 'cognitive psychometric models' to refer to the collective body of models concerned with the parameterisation of item-level or person-level factors associated with test performance. However, this terminology is not commonplace within the literature, and it is far more common to see reference to 'cognitive diagnostic models', and to a lesser extent, 'diagnostic classification models'. It is therefore necessary to explain why I have adopted this less common terminology for this thesis.

Nichols (1994) first coined the term ‘cognitively diagnostic assessment’, now more commonly referred to as ‘*cognitive* diagnostic assessment’ (CDA), to refer to assessments designed to allow for more granular interpretations of examinee performance driven by insights from cognitive psychology. According to Nichols, CDA includes the use of problems, tasks or items designed through systematic research-based variations in the characteristics of what is assessed, and are scored in ways that allow for inferences about the different qualities of students’ knowledge and cognitive processes. He also contrasted CDA with assessment derived from classical test theory and assessments from item-response theory backgrounds, noting that these approaches typically place greater emphasis on scores in content areas, rather than on examinee’s cognitive processes.

Rupp (2007) argued that the ‘diagnostic’ part of this term could sometimes be limiting, because several strands of relevant research, such as those using cognitive psychology to inform automated item generation, were not directly concerned with any form of ‘diagnosis’. In other words, the ‘diagnosis’ part of the term is less relevant to research concerned with item-level parameterisation and explanation. Consequently, Rupp proposed the terminology of ‘cognitive *psychometric* models’ to broaden the scope of relevant models beyond those concerned with explaining examinee performance but that are still rooted from insights from cognitive psychology.

It is worth noting that this terminology has not been commonly adopted since. Indeed, in his more recent publications, Rupp has typically adopted the cognitive *diagnostic* terminology that is more consistent with other published research since his proposal. Nonetheless, for the purpose of this thesis, I have chosen to adopt the cognitive *psychometric* term, as I believe it best encapsulates one of the central ideas of the thesis. That is, that the item-level and person-level parameterisation should be seen as complementary, and that item-level parameterisation (which is not directly concerned with measuring or classifying the abilities of examinees) functions as an essential precursor to later diagnostic analysis. The thesis will therefore refer primarily to cognitive *psychometric* models. However, where appropriate, the discussion may instead refer to

cognitive *diagnostic* models when talking about the collective body of literature/models that focus exclusively on person-focused diagnoses. This is particularly the case in paper 3 of this thesis.

2.2.2 What are cognitive psychometric models?

Defining what 'counts' as a cognitive psychometric model (CPM) is not as straightforward a task as one might hope. This is, in part, due to the rise of models from varying statistical and philosophical traditions with shared characteristics and/or goals. In part, the choice to adopt the term 'cognitive *psychometric* models' arises from a desire to unite these differing strands under a common umbrella, though I cannot promise that proponents from these varying traditions would agree on this language choice, or indeed, my argument for a need for an umbrella term at all. Despite some potential for dissent, the main models considered under this CPM umbrella and used throughout this thesis have previously been considered as related (albeit, sometimes distantly; e.g. Williamson, 2023; Zhang et al., 2023) and I believe, should be seen as complementary.

CPMs can be thought of as statistical approaches that parameterise responses to test-items into specific components in ways that are guided by cognitive psychology and/or pedagogy. In most CPMs, parameters are defined by content experts in a Q-matrix, otherwise known as an item-by-attribute matrix, in which experts stipulate which properties of test-items are expected or anticipated to affect the probability of an examinee giving a correct or desired response. An example Q-matrix was presented in section 1.2. In the context of an educational assessment, these properties might include content areas hypothesised to relate to item difficulty, the different kinds of cognitive processes demanded by the item, and other aspects of how the test-item is presented or administered which may affect the ways in which examinees interact with test-items. Potential sources of item difficulty, particularly in the context of mathematics assessment, are discussed further in section 2.3.

A core assumption of CPMs is that examinees' responses to test-items tell us something about either (a) properties of the items, (b) properties of the examinees, or (c) simultaneously provide information on both. Most applications of CPMs will acknowledge that complex interactions exist

between examinees and items. However, to draw the most detailed interpretations of either examinees or items, most CPMs focus on drawing stronger inferences on one of these and make simpler assumptions about the nature of the other. For example, the linear logistic test-model (LLTM; Fischer, 1972) is an example of a CPM that looks to explain item difficulties based on the content and properties of those items, but still estimates person abilities unidimensionally. By contrast, person-diagnostic CPMs such as the rule-space method (Tatsuoka, 1983) or the deterministic noisy inputs 'and' gate model (DINA; de la Torre, 2009) look to describe differences in examinee performance based on patterns of mastery and non-mastery of skills required by test-items and therefore focus their interpretations on what test performance data says about individual examinees, with little to no focus on item evaluation.

2.2.3 Item-response theory models and the Rasch model

The one-parameter logistic model (1PL) and its related family of models (e.g. 2PL, 3PL, and variants applied for polytomously scored items) are among the most widely used IRT models in scoring and scaling of assessment data. This includes in international assessments such as TIMSS, PISA, and PIRLS, which use these models to estimate psychometric properties of test-items and from this, derive the plausible values discussed in section 2.1.1.

The 1PL focuses on a single item-level parameter, item difficulty (β), and a single person-level parameter, latent ability (θ). These estimates are independent of each other in that, unlike in the item or person scoring estimates produced under classical test theory models, the estimates of item difficulty produced by IRT models do not depend on the ability of the examinees. Item difficulty estimates and examinee ability estimates are however calculated on the same scale, and allow for estimations of the probability of an examinee's success on a test-item as a function of the estimate of that examinee's ability relative to the difficulty of the item on that scale.

The formula for the 1PL model is as follows:

$$P\left(X_{ie}=1 \vee \theta_e, \beta_i\right)=\frac{\exp^a\left(\theta_e-\beta_i\right)}{1+\exp^a\left(\theta_e-\beta_i\right)}$$

The probability of an examinee e providing a correct response to test-item $i\ (X_{ie}=1)$ is a function of the examinee's latent ability (θ_e) and the difficulty of the item (β_i). An item's estimated difficulty is equal to the estimated ability level of an examinee who has a 50% chance of providing the correct response. Thus, if an examinee's estimated ability level is greater than the estimated difficulty of a given item, the model suggests that there is a greater than 50% chance that the examinee provides the correct response. Under the assumptions of the 1PL, all items are equally discriminating (α) – that is, they are all equally capable of discerning between higher ability and lower ability examinees. The 2PL variant of this model changes the equation such that individual items can be more or less discriminating than others (a_i). The 3PL meanwhile introduces a pseudo-guessing parameter, which accounts for the estimated probability of correct guessing from lower ability examinees associated with each item. Item parameter estimation in TIMSS 2011 was conducted through the use of both 2PL and 3PL models, for constructed-response and multiple-choice items respectively.

From a purely mathematical perspective, the 1PL can be seen as a slightly more general case of the Rasch model (Rasch, 1960). The only difference, which has little direct impact on the estimation of item or person parameters, is that the Rasch model requires that all items are equally discriminating (equivalent of a in the 1PL = 1 for all items), while the 1PL assumes this value is equal for all items, but not necessarily equal to 1. However, the theoretical assumptions of the models and the interpretations about what the results of the two models tell us about the data are fundamentally different. This is because, in essence, a Rasch analysis looks to assess how well the data fits the Rasch model, whereas IRT models such as the 1PL look to assess how well the 1PL can explain the data. Evidence of a poor relationship between model and data and therefore interpreted very differently depending on which theory of measurement is followed.

Proponents of the Rasch model would likely argue that items which fit the Rasch model poorly are indicative of an issue with the item, rather than the model (Andrich, 2004). One of the philosophical positions underlying the Rasch model is that it should be possible to measure an examinee's ability on a linear, invariant scale (Rasch, 1960). Items which deviate strongly from

this may be assumed to assess a construct other than that the unidimensional construct that the rest of the test is measuring, or may be otherwise poorly constructed. In other words, these items may violate the assumptions required for invariant measurement. A proponent of the Rasch model over IRT models with more parameters may therefore see poor item fit as grounds for item revision and/or exclusion. By contrast, the position underlying many IRT models, including the 1PL, is that poor item fit to the model is more likely to arise because the model does not adequately account for the complexity of the data (Linacre, 2005). Alternative model selection and the introduction of additional item-level parameters (e.g. through the 2PL and/or 3PL models) can instead be used to improve fit of the model to the data to better account for the data's complexity, even if this comes at the expense of invariant measurement. Such additional parameters would not be added to the Rasch model, because invariance is seen as a necessary, and not just preferred, characteristic of item and person measurement.

In this thesis, and particularly in paper 2, I have chosen to adopt the 1PL terminology. This decision was made primarily to facilitate drawing comparisons to the item-level parameterisation conducted by the IEA for TIMSS test-items. However, it is not the case that I believe the greater flexibility of the 2PL/3PL models is necessarily a strength or advantage of these models over the Rasch model.

An assumption of the 1PL is that estimates of the difficulty of a set of items in a test calculated under the 1PL should be the same (within a small standard error of measurement) if administered to a set of higher performing examinees as they would be if administered to a set of lower performing examinees. Similarly, the estimates of each examinee's ability should not be dependent on the difficulty of the test-items they were given (again within a small standard error of measurement). Therefore, two examinees can be given entirely different test-items but can still be directly compared in terms of their ability level, assuming there is other information that allows the two sets of test-items to be compared on a single scale. This invariance in the model enables large-scale international assessments such as TIMSS to administer a much wider set of test-items than might otherwise be possible, allowing for more stable parameter estimation that can mitigate some item-level noise.

The 1PL model is sometimes classified as a descriptive or doubly descriptive model (e.g. De Boeck & Wilson, 2004), because it describes the variation in person abilities through a person-level parameter (θ_e) and also describes the variation in item difficulties through an item-level parameter (β_i). Other models, however, look to *explain* either of these parameters with respect to item or person properties. A latent regression Rasch model, for example, can be used to explain the person parameter with respect to a given set of person properties (e.g. gender, age, country, prior attainment). Person characteristics can therefore be used to explain differences in performance. It should be noted that this differs from the person explanatory elements in other models derived from cognitive diagnostic traditions, where item-level performance is used to 'explain' what students know. Of greater concern to this thesis are the item-explanatory models which look to explain item difficulties with respect to characteristics of the items and what they assess.

2.2.4 Item explanatory models – the linear logistic test model

One of the earliest examples of item explanatory models was the linear logistic test model (LLTM), initially proposed by Fischer (1972). The LLTM can be applied to response data for dichotomously scored test-items, and provides an extension to the calculation of item difficulty parameters in the 1PL model. While the LLTM still assumes unidimensionality of the test, it also assumes that the difficulty of each test-item is the composite of various different factors associated with that construct, or as Fischer referred to them, 'sources of cognitive complexity'. In the case of a mathematics assessment, for example, the LLTM assumes that the test as a whole assesses a singular construct – mathematical ability – but that each item of the test is made more or less difficult by the different facets of mathematics that the item tests, or by other features of those test-items.

The LLTM imposes a linear constraint on the calculation of item difficulty parameters (β_i) in the formula for the 1PL, such that:

$$\beta_i = \sum_{j=1}^{J} q_{ij} n_j + c$$

where q_{ij} is the weight of each proposed difficulty factor j for item i, and n_j is the estimated difficulty parameter for each proposed difficulty factor / attribute j. A normalisation constant, c, is typically introduced, such that the mean of item difficulties under the LLTM is equal to 0. This facilitates comparisons with other mean-adjusted models, including the 1PL where the mean item difficulty is typically set to equal 0.

Embretson and Daniel (2008) argued that the LLTM is a useful method for establishing the validity of constructs used to design test-items, and allows for predictions to be made about the difficulty of other test-items. This sort of validation has the potential for wide ranging benefits to item development and test administration, including in the development of item banks for large-scale assessments and in automated item generation. These benefits are of course reliant on the LLTM being able to accurately predict item difficulties from the accurate selection and specification of attributes in the Q-matrix, and that item difficulty does indeed summate linearly without other more complex interactions between sources of item difficulty.

It should be noted that the formula for the LLTM does not include an error term. The LLTM can be seen as a highly restrictive model (Wilson et al., 2008) because it makes a very strong assumption about the exact additive contribution of the parameters associated with test-item factors (or, in Fischer's terminology, the item-level sources of cognitive complexity) to item-level difficulties. In practice, this assumption is rarely met in real-world datasets (Hartig et al., 2012), and linear logistic test modelling of item difficulties typically produce poor model fit relative to logistic models such as the 1PL, 2PL, or 3PL. One approach to compensate for this limitation is to instead use a linear logistic test model with random item effects (LLTM+e), as presented by Janssen et al. (2000, 2004). While this model relaxes the strict assumptions made by the LLTM, one major practical limitation of the LLTM+e are the heavily increased computational demands, particularly on larger assessment datasets with large numbers of items and/or persons. Further, a major strength of the LLTM is the ease of interpreting results. By contrast, interpretation becomes more complex in the LLTM+e due to the addition of random effects. The LLTM+e has therefore seen rather limited use outside of smaller scale simulation studies.

A common approach to exploring goodness-of-fit of the LLTM is to first compare LLTM fit values to the equivalent values of the 1PL or Rasch model (Baghaei & Kubinger, 2015). Three values - deviance, Akaike's information criterion (AIC; Akaike, 1974), and Bayesian information criterion (BIC; Schwarz, 1978) are typically reported. Deviance is calculated as the negative of 2 log-likelihoods of the maximum likelihood, given the model (Wilson et al., 2008). Both the AIC and BIC are derived from this deviance value but apply different penalties based on the number of parameters. Comparisons of model fit are discussed further in section 3.3.3.

2.2.5 Person-diagnostic models

There are also CPMs derived from Rasch and IRT backgrounds that look to explain differences in item-level performance based on person characteristics, rather than item properties. Person *explanatory* models such as the latent regression Rasch model look to explain differences in item-level performance with respect to person characteristics. This is done by regressing person ability estimates onto examinee characteristics; this can include both categorical characteristics such as gender, or continuous characteristics such as prior attainment scores. However, this thesis is more concerned with 'person-*diagnostic*' models, which instead use item-level performance data to make diagnostic inferences beyond single latent ability estimates. These are often referred to as cognitive diagnostic, or diagnostic classification models, but here, I use the term 'person-diagnostic' to refer to these models which place emphasis on describing examinee characteristics in order to draw parallels with the 'item-explanatory' models discussed previously.

Zhang et al. (2023) suggested that the body of diagnostic models, while arising from a variety of theoretical and applied contexts, share a unifying set of characteristics and assumptions. The most defining of these is the assumption that test performance depends on the examinees' 'mastery' (Tatsuoka, 2009) of a set of pre-defined attributes (skills, procedures, and knowledge) assessed by the test-items. These attributes, depending on the model, can be either entirely separate, related with unknown structure, or seen as hierarchical (i.e. mastery of one attribute is at least partially dependent upon the mastery of one or more others in an a priori described manner). As with the LLTM in item-explanatory modelling, for example, the assignment of

attributes to items is almost always made through the use of a Q-matrix. While not aiming to achieve the same goals, there are many similarities in the approaches adopted by the LLTM and these kinds of person-diagnostic models.

One of the main families of person-diagnostic models are those within the deterministic inputs, noisy ‘and’ gate (DINA) family (de la Torre, 2009, de la Torre, 2012). While relatively new, they have become among the most widely used person-diagnostic models in the last 15 years (Williamson, 2023) in educational research. These are latent class models that categorise examinees into patterns of attribute mastery based on item-level performance.

The underlying assumption of all the models within the DINA family is that an examinee’s success on an individual test-item is a function of their ‘mastery’ of the attributes that the item assesses. That is, whether or not the examinee has a sufficient understanding and/or proficiency with the attribute assessed by that item. These attributes may relate to subject content, or cognitive processes and skills that are evoked by the test-item. The different models within this ‘DINA family’ vary in their assumptions about the interactions between attributes and the necessity of mastery of all attributes for success. For example, the original DINA model (de la Torre, 2009) assumes that examinees should provide a correct answer to a given question if they have mastered *all* of the attributes associated with an item, and provide an incorrect response to the item if they have not mastered all attributes. The DINA model is considered a conjunctive, or non-compensatory model, in that mastery of *all* associated attributes is modelled as necessary for success on a given item, and no level of mastery of the other attributes can compensate for a lack of mastery in one. This is reflected in the ‘and’ gate naming of the DINA model.

The DINA model does somewhat account for systematic noise introduced by examinee behaviour (and hence the gate is labelled as ‘noisy’). If an examinee provides a correct response despite non-mastery of all the required attributes, the interpretation is that the examinee has guessed the correct answer. Conversely, it is possible for an examinee to ‘slip’ on an item, which occurs when the examinee has mastery of all required attributes yet fails to provide the correct response. Two parameters, ‘guessing’ and ‘slipping’ are calculated for each item.

One of the main advantages of the original DINA model is in its ease of interpretation. It has been argued that the DINA model is a viable diagnostic model for practical application outside of research exercises because it produces cognitive profiles of students that are easily interpretable and that could be applied in real-world contexts to inform and improve students' learning (e.g. de la Torre, 2009, 2012; Lee et al.., 2011). However, more recent studies comparing CDMs often find that the original DINA model is typically not the preferred model at either the item or test-level (e.g. Delafontaine, 2022; Williamson, 2023). The DINA model typically produces high proportions of attribute mastery among students, and individual items often have very high item-level guess and/or slip parameters (suggesting poor model fit). Perhaps most importantly, the DINA model has been shown to fit many real assessment datasets poorly, including in TIMSS compared to alternative, less parsimonious CDMs (e.g. Yamaguchi & Okada, 2018; Ma & de la Torre, 2020a). A similar issue is shared with the equally parsimonious though less common 'DINO' model (deterministic inputs, noisy 'or' gate), which assumes mastery of just a single attribute is sufficient for success, and there is no further benefit associated with the mastery of more than one attribute. The DINO model has been reported to fit large assessment datasets equally or even less poorly than the DINA (e.g. Yamaguchi & Okada, 2018).

de la Torre (2012) later proposed a generalized DINA (G-DINA) model which relaxed some of the assumptions about the probability of success associated with each attribute. Rather than calculating just two parameters for each item (guessing and slip), the G-DINA calculates 2^J parameters, where J represents the number of attributes associated with the item. The G-DINA calculates separate guess and slip parameters for each potential pattern of attribute mastery. So, for example, an item with three associated attributes would have 8 calculated parameters, representing each of the possible combinations of attribute mastery and non-mastery. The G-DINA therefore differs from the DINA (or DINO) models because it does not make a-priori assumptions about the relationships of attributes.

The G-DINA model has become substantially more popular in CDM research using data from ILSAs in recent years because it is often found to be better fitting among the DINA family of models than the DINA or DINO data (de la Torre, 2012; Ravand & Baghaei, 2020; Williamson,

2023; Yamaguchi & Okada, 2018). The main weakness of these more generalised models relative to the DINA model is arguably in the ease of interpretation, as the larger number of item parameters makes interpretation substantially more convoluted and less functionally applicable (de la Torre, 2012; Williamson, 2023). As is the case with item-explanatory CPMs, there is a trade-off with general and specific person-diagnostic models (Deonovic et al., 2019), with the specific models being easier to interpret and typically requiring smaller sizes, while more generalised models produce greater model fit and require less assumptions. Ravand and Baghaei (2020) argue that, as a rule of thumb, generalised person-diagnostic models are preferred when sample sizes suffice (more than 5,000 examinees), while using more specific, saturated models such as the DINA may be more appropriate when sample sizes are smaller, or when the Q-matrix includes more than five attributes.

One of the main critiques of these analyses is in the assumptions they make about the mapping of attributes to items, both in terms of a) the accurate specification of attributes to items (i.e. are the attributes tested by items correctly listed within the Q-matrix used for analysis?), and b) even if the attributes are correctly specified, do they in any way actually relate to / impact performance? In other words, do the attributes associated with the item have anything to do with the difficulties of those items? For example, von Davier (2018) argued that, while being conceptually attractive, there are several concerning assumptions associated with these person-diagnostic models, including the assumptions that skills can be represented with two possible outcomes (mastery or non-mastery) and that the interaction of attributes can accurately be determined (i.e. additive, multiplicative, compensatory vs. non-compensatory etc.).

We cannot assume that the success of student *A* and failure of student *B* on a given item tells us about their specific strengths and weaknesses (or, ‘mastery’) if we do not fully understand exactly what the item tests, or how examinees experience those items. If we can explain that item *X* is more difficult than item *Y* because of specific factors associated with those items, be they subject-specific content areas, the cognitive processes evoked by those items, or other factors related to their administration (e.g. test language, response modes, time constraints), then we can make stronger inferences about what that tells us about how examinees engaged with those

items, and what that ultimately implies about their knowledge states. As the level of specificity in the claims made about an examinee's knowledge and abilities increases, so too must the empirical support for how those specific claims were derived from the data produced by the measures (items). This is not something that the vast majority of person-diagnostic analyses are able to offer, particularly in retrofitting analyses where the attributes used for diagnosis might not be explicitly tied to individual items. This therefore undermines the validity of the conclusions drawn by these analyses.

A central argument to this thesis is that item explanatory modelling is not only complementary to person-diagnostic modelling, but a necessary precursor. This necessary precursor is often neglected by person-diagnostic modelling studies. Establishing what item-level factors and properties contribute to the difficulty of test-items in educational assessments, both broadly and in specific assessments such as TIMSS mathematics is therefore required.

2.3 Item difficulty

In their opening discussion of what factors contribute to the difficulty of items in GCSE mathematics examinations, Fisher-Hoch and Hughes (1996) cited a comment from Stenner (1978), in which he suggested that if a test-item developer does not understand "why this question is harder than [*another question*], then you don't know what you're measuring". Fisher-Hoch and Hughes interpreted Stenner's comment as raising a concern about the construct validity of these assessments, and subsequently, additional concerns about the conclusions drawn about what the examinees of those assessments know and can do. While their discussion focused on these threats in the context of national GCSE assessments in England, construct validity of test-items must be a primary concern of all assessment, particularly those whose conclusions yield powerful or wide impacts, be they political or pedagogic, or if they have direct impacts and outcomes on those who are examined.

Yet predicting the difficulty of test-items prior to their administration remains a challenge, even for experienced test-item developers and for test material that may, at first glance, appear simplistic in content or design (Stiller et al., 2016). One advantage of studies such as TIMSS is that the

items go through thorough pre-testing, and so there is, in theory, greater potential scope to assess item difficulties (and identify poorly functioning items) prior to their main test administration. However, the slow progress that has been made in the accurate prediction of the difficulties of real-world test-items was lamented by El Masri et al. (2017), who attributed this to "poor theoretical models of item difficulty" (p.78), and noted the consequences this has on the ability to draw valid inferences about students' understanding based on assessment data.

Before discussing specific item-level factors that have been predicted to influence item difficulty, it is worth clarifying exactly what is meant by 'item difficulty'. One need for this clarification is because it has been suggested that the term 'item difficulty' has been used to describe both an intrinsic psychometric property of a test-item, and also to describe the interaction between an item and how it is experienced by examinees and their rate of successful response (e.g. Pollitt et al., 2007).

Pollitt et al. (2007) argued that a distinction should be drawn between the difficulty of a test-item and the demands of an item. Under this conceptualisation, the *demands* of an item relate to the cognitive processes required to produce a correct response. The *difficulty* of an item meanwhile is a more purely psychometric term relating to the position of the item on a scale ranging from the item with the greatest likelihood of a correct response to the item with the lowest likelihood. While there is a strong overlap between the items that are most demanding and those that are the most difficult, other factors not directly associated with the demands of the item can affect difficulty.

In the context of international assessments such as TIMSS, the estimated difficulty of items can vary quite widely when using data from national samples, even among those sharing the same test language. Students' experiences with different national curricula, for example, could explain differences in subscale scores between countries. In TIMSS 2023, students in England scored significantly higher on the data domain than the other content domains, with significantly lower scores in measurement and geometry. The authors of the TIMSS 2023 national report for England (Golding et al., 2024) commented that this had been a consistent pattern in England since 2015 (and to a lesser extent in most other Anglophone countries), but also noted that this

contrasted with the patterns observed in the majority of other high-performing countries, particularly those in East Asia (Bokhove et al., 2022).

However, under Pollitt's definition, the inherent demands of the item are the same regardless of who the examinees are, what they know, and what they can do; it is the interaction of the examinees with the item that affects item difficulty. These definitions may therefore appear to be at odds with the assumptions of some cognitive psychometric models that consider item difficulty to be a property of an item that is invariant with respect to the ability of the examinee (e.g. the Rasch / 1PL model). If we were to attempt to empirically estimate the difficulty of the test-items used in TIMSS based solely on England's cohort, then we would probably find that the difficulty estimates for measurement and geometry items were higher than if we had used data from students in a country with relative high measurement and geometry scores. Thus, disentangling the concepts of difficulty and demands can be challenging, and arguably contributes to the complexities in modelling item difficulty empirically.

2.3.1 Sources of item difficulty

Pollitt et al. (1985) proposed that item difficulty in educational assessments are driven by three main categories of difficulty. The first of these is concept difficulty; that is, that the subject matter of some test-items is intrinsically more complex than others. We might otherwise refer to this as difficulty arising from test-item content. However, Pollitt et al. (2007) later drew more explicit lines between difficulty and demands, relating the latter to how examinees experience items, with content familiarity being a key driver of item demands. However, some topics are fundamentally more abstract or detailed than others, and following Pollitt et al.'s (1985) argument, this a key driver of item *difficulty.*

Secondly, items have a process difficulty, which relates to the kinds of cognitive operations that the items evoke, and how the items put demands on working memory. Some test-items may ask examinees to explain ideas, draw generalisations, or extract/apply principles to data. Test-items may also vary in how many different processes are simultaneously required (e.g. single vs. multi-

step problems). The types and quantities of cognitive process would all have unique and interacting effects on item difficulty.

Finally, Pollitt et al. (2007) suggested items have a ‘question’ difficulty, or what we might otherwise refer to as a presentational difficulty, that arises from how the item is physically presented to the examinee through language, images, and how/where it may appear in relation to other test-items. This can also relate to the type of response expected of an examinee, such as a multiple-choice or open-ended response, or how the item is presented to help guide or distract the examinee towards expected responses.

Evaluating test-items to establish the content and cognitive processes they assess is, as previously discussed, more challenging than might be expected. This in turn, contributes to the challenges in modelling item difficulties. While we can attempt to analyse items and predict what content examinees will find challenging, and what cognitive processes they will require in order to reach their answers, empirical data of test-item functioning often suggests that they can often function unpredictably because examinees can, for example, adopt unexpected approaches to reach correct answers, or focus on irrelevant information that is not intended to drive item difficulty.

Both qualitative and quantitative approaches have been used to investigate item functioning and assess item difficulty. For example, Fisher-Hoch and Hughes (1996), adopting a primarily qualitative approach, investigated the errors made by students who participated in the examinations of the 1994 ‘School Mathematics Project’, which closely mirrored the content covered in the 1994 set of GCSE examinations in mathematics. Test-items were distributed across six papers, with each paper designed to represent an increase in difficulty. Students were administered two adjacent papers (e.g. papers 3 and 4). Focusing on a cohort of 600 scripts (100 from each paper), Fisher-Hoch and Hughes identified the errors made in each script, and then sorted the scripts into groups based on common errors. They then drafted a list of what they believed were the common sources of difficulty in these items. Their proposed list is presented in Table 2.

Table 2: Proposed sources of difficulty in mathematics test (adapted from Fisher-Hoch & Hughes, 1996)

Stage of question engagement	Source of difficulty	Description
Question recognition The examinee has to determine whether they can attempt the question, and whether they have the required set of knowledge and skills.	Command words	Inconsistencies in how words such as 'explain', 'find', 'state' etc. are used in question stems may increase item difficulties.
	Context	Scenarios in some questions have been inaccessible / abstract to some examinees, which may have limited the examinees' abilities to construct mental models of the question.
	Stated principle	Questions which do not explicitly state what the topic or concept being examined is may increase difficulty.
	Combination of topics	Questions that involve more than one mathematical topic are likely harder (note: it is not explicitly stated how they anticipate the combination to interact)
	Isolated skill or knowledge	The question covers a less-well represented topic in the syllabus.
Understanding the problem The examinee has to create a mental model of how to solve the problem.	Mathematical language	The question requires an understanding of the meaning of mathematically relevant terminology.
	Mathematical vs. everyday language	The question includes words that have both a specific meaning in mathematics, and at least one other meaning in an everyday context (e.g. mean, range).
	Mathematical sequencing	Sub-parts of questions require backtracking through the item stem.
Planning The examinee has to identify what strategies they can use, and what information/data they require.	Recall strategy	If the strategy required is not explicitly stated, then the item is likely to be more difficult.
	Alternative strategies	If the examinee adopts a strategy that is not expected by the test-item writers, the difficulty of the item may be considerably greater (or smaller) than anticipated.
	Abstraction required	Items that require some level of abstract thinking are likely to be of higher difficulty than those that do not.
	Spatial representation required	Questions which require spatial reasoning to build a mental model of the question are likely to be more difficult.
Extraction The examinee has to locate the information that is required	Ambiguous resources	Unclear diagrams can increase item difficulty.
	Irrelevant information	Irrelevant information can be distracting and increase item difficulty.
Execution The examinee has to calculate or produce the desired/correct response	Number of steps	Questions with a large number of steps may overload working memory, causing the examinee to lose important information and therefore increase item difficulty.
	Arithmetic errors	Some questions provided greater opportunities for simple arithmetic mistakes, therefore increasing estimates of item difficulty.
Recording The examinee has to write down their solution.	Paper layout	Physical organisation of the question, including question ordering, may increase (or decrease) item difficulties.

Rather than focusing on specific mathematical topics and their relations to item difficulties, Fisher-Hoch and Hughes' list of sources of item difficulties focus on the processes involved in how examinees engage with mathematics items. An example item included in their analysis is presented in Figure 2. This item, from paper 1 (the easiest of the six papers) asked examinees to identify an acute angle and a reflex angle, from the given shape. The most common response, among both higher-performing and lower-performing examinees who were given this item, was to incorrectly identify the exterior obtuse angle as the reflex angle. The second most common response was to identify the interior obtuse angle as being a reflex angle.

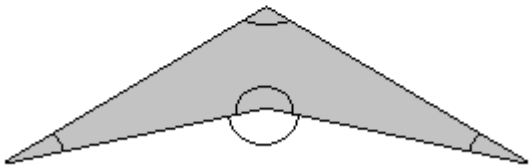
Figure 2: Example base item adapted from Fisher-Hoch and Hughes (1996)

This shape is called an arrow head.

Mark and label **clearly**

(a) an acute angle, [1]

(b) a reflex angle. [1]



Three sources of difficulty from Table 2 were associated with this item.

- Context – the opening statement, 'this shape is called an arrow head' provides additional information that is not required. Fisher-Hoch and Hughes referred to this as an 'invalid' source of difficulty.
- Recall strategy – the examinee needs to recall what the different kinds of angles are. This was identified as a 'valid' (i.e. mathematically relevant source of difficulty).
- Ambiguous resources – the four interior angles and one exterior angle are already marked on the diagram. Fisher-Hoch and Hughes noted that many examinees had appeared to struggle with how to 'mark clearly' on the diagram where angle arcs had already been drawn, particularly at the point where the obtuse and reflex angles meet. This potentially ambiguous information was marked as an invalid source of item difficulty.

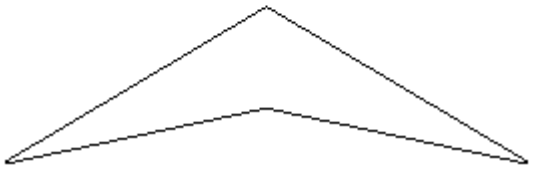
In a subsequent paper (Fisher-Hoch et al., 1997), they presented two variants of the same item in which the suggested invalid sources of item difficulty had been reduced or removed. This was done by removing the unnecessary contextual statement, and removing angle arcs from the diagram. Figure 3 shows the same item as in Figure 2 but with both of these recommended changes. Note that for their own further research purposes, they proposed presenting variants of this item with only one of these changes implemented. This was to assess how each change independently may contribute to changes in item difficulty.

Figure 3: Example revised item adapted from Fisher-Hoch and Hughes (1996)

Mark and label **clearly**

(a) an acute angle, [1]

(b) a reflex angle. [1]



Some changes were also made to what were considered a 'valid' sources of item difficulty. Figure 4 shows a test-item before and after revisions. In addition to the reformatting of the numerical sequence from a list to a table, they also changed a sub-item, part (b), from a calculation of the 25^{th} term to the 10^{th} term in the sequence. This was because their qualitative analysis of errors showed that many students who took the original item reached incorrect final answers because of arithmetic errors made in continuing the sequence from the 4^{th} to the 5^{th} to the 6^{th} terms etc. all the way up to the 25^{th} term, rather than working out the n^{th} term and substituting in the required value of n. Reducing the required term here was done to reduce the chance of students making arithmetic errors. However, not all the implemented changes to items successfully altered item difficulties, and in some cases, produced the opposite of the intended effect. For example, while additional contextual information was sometimes found to be unnecessary, as shown in Figure 4, in other items, particularly those relating to data handling, they found that the removal of

contextual information, even if not directly necessary for reaching the solution, sometimes ended up causing the item to be more difficult than the original version.

Figure 4: Example revised item adapted from Fisher-Hoch et al. (1997)

Look at the sequence

2 8 18 32

(a) Write down the next term of the sequence.

(b) Write down the 25th term of the sequence.

(c) Write down the *n*th term of the sequence.

Look at the sequence

Position	*Term*
1	2
2	8
3	18
4	32
n	?

(a) Write down the next term of the sequence.

(b) Write down the 10th term of the sequence.

(c) Write down the *n*th term of the sequence.

2.3.2 Reading-related contributions to item difficulty

Other studies have tried to empirically assess mathematical content and cognitive complexity contributions to item difficulty with limited success, with other item-level features such as language often having stronger relations to item difficulty estimates. For example, Ferrara et al. (2011) utilised data from a state mathematics assessment in the USA aimed at students in grades 3, 4 and 5 to explore how item-level performance compared to proficiency level

descriptors used in the assessment at each curriculum grade. In modelling the difficulty of items, they drew from a wide range of potential contributing factors, including the state assessment's own learning progression within specific mathematical content areas (e.g. number and operations, geometry etc.), a general measure of reading load, 'depth of knowledge' (Webb, 2007), a simple three-category framework of mathematical complexity derived from the National Assessment of Educational Progress (NAEP) assessment in the USA (low/moderate/high complexity), a simple three-category framework of question types taken from Ferrara's previous work in science assessment (Ferrara & Duncan, 2011), and finally, a condensed version of a readability framework developed by Shaftel et al. (2006) for mathematics test-items. Each item was coded by a team of raters, and then entered as independent variables into a linear regression onto item difficulty estimates. Most of these factors did not significantly contribute to the amount of variance in item difficulty estimates. Reading load, and aspects of Shaftel et al.'s readability framework, including ambiguous vocabulary or technical mathematical vocabulary, were among the few significant predictors of item difficulty.

Other studies of mathematics word-problems (Daroczy et al., 2015), NAEP mathematics (Abedi & Lord, 2001), items used in national versions of TIMSS mathematics at the eighth grade (Nyström, 2008), as well as items used in TIMSS science (Rosca, 2004; Glynn, 2012) and key stage 2 science assessments (El Masri et al., 2017) have all suggested that test-item language, including both subject/context relevant terminology and unintended sources of difficulty arising from irrelevant/confusing language and/or overall reading load have significant effects on item difficulty.

3 Methods

Section 3.1 focuses on the methods underlying paper 1, which primarily addresses research sub-question 1. Section 3.2 provides a shared discussion of the data and models used in papers 2 and 3, as there is substantial overlap between these. Sections 3.3 and 3.4 provide focused discussions of more specific methodological choices and approaches in paper 2 and paper 3 respectively. These two papers are targeted to research sub-questions 2 and 3 respectively.

3.1 Paper 1 – thematic review

Paper 1 focuses on addressing **research sub-question 1** - ***What are the main themes, foci, and findings of research utilising international assessment data for cognitive psychometric modelling purposes?*** To address this research question, a thematic review was conducted, targeting four digital databases; ERIC, Scopus, PsycInfo, and Web of Science. These databases were chosen based on prior knowledge of their excellent coverage of studies concerning international large-scale assessment. While the focus of the thesis as a whole is on TIMSS, the similarities to PISA and PIRLS warranted their inclusion, and I wanted to assess how commonly TIMSS was used for these kinds of analysis relative to other subject areas.

Previous review articles have focused more broadly on cognitive diagnostic models (e.g. Williamson, 2023), or on research outputs concerned with specific international assessments (e.g. Hopfenbeck et al., 2018; Stiff et al., 2023; Cordero et al., 2017). To my knowledge however, there have been no previous thematic or systematic reviews that have attempted to synthesise the findings of research using ILSA data to conduct CPM analyses.

In this thesis and in the paper itself, I have chosen to refer to this as a *thematic* review, rather than a *systematic* review, which had been my original intention for the paper. The general approach to this review article *was* conducted in accordance with the PRISMA 2020 guidelines for systematic reviews (PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses). This lists 27 recommendations of systematic review content that facilitates transparent reporting. This includes, but is not limited to, explicit reporting of eligibility criteria, search

strategies, and an assessment of bias in the located articles, as well as in the interpretation of main findings and the synthesis process. However, on reflection, referring to this paper as a systematic review felt inappropriate. This was because the research sub-question for this paper is relatively open and broad, whereas a traditional systematic review would likely address a more targeted research question. This breadth is reflected in the wide range of search terms included (as outlined in section 3.1.1). As the goal of the article was to identify common *themes* amongst the published literature that had applied CPMs to ILSA data, I settled on the term 'thematic review'.

3.1.1 Search terms

A list of search terms were devised based on areas of interest and prior readings of CPM analyses of TIMSS data and of relevant theoretical and methodological literature. Searches were made in all fields (title, abstract, keywords etc.) using a combination of two terms from a pair of lists, as reported in Table 3. Each search was made using one of the three international large-scale assessments in List A, and one of the 27 statistical models or approaches in List B. This led to a total of 81 searches per database. The search terms in List B were chosen to cover both a wide range of CPMs employed in item difficulty modelling and cognitive diagnostic analyses, as well as other terminology common to the broader scope of difficulty modelling literature.

Table 3: List of search terms used in Paper 1's thematic review

List A search terms (ILSA studies)
One of:
‘‘PISA" OR ‘‘Programme for International Student Assessment"
‘‘TIMSS" OR ‘‘Trends in International Mathematics and Science Study"
‘‘PIRLS" OR ‘‘Progress in International Reading Literacy Study"
List B search terms (models / statistical approaches)
One of:
“item difficulty”
‘‘response demand"
‘‘item demand"
‘‘cognitive demand"
‘‘text demand"
“cognitive complexity”
‘‘text complexity"
‘‘cognitive analysis"
‘‘diagnostic assessment"
‘‘item model*ing"
‘‘response model*ing"
‘‘CDM" OR ‘‘cognitive diagnostic model*ing"
‘‘CPM" OR ‘‘cognitive psychometric model*ing"
‘‘DINA" OR ‘‘DINO" OR ‘‘deterministic input noisy"
‘‘RSM" OR ‘‘rule space method" OR ‘‘rule-space method"
‘‘RUM" OR ‘‘unified model" OR ‘‘reparameteri*ed unified model"
‘‘Fusion model"
‘‘LLTM" OR ‘‘linear logistic test model"
‘‘LPCM" OR ‘‘linear partial credit model"
‘‘automated item generation"
‘‘item characteristics"
‘‘design matrix"
‘‘Q-matrix" OR ‘‘Q matrix"
‘‘evidence cent*red design" OR ‘‘evidence-cent*red design"
‘‘DIF" OR ‘‘differential item functioning"
‘‘DTF" OR ‘‘differential text functioning"
‘‘DBF" OR ‘‘differential bundle functioning"

In order to focus only on articles of relevance to the research sub-question, articles needed to include data from at least one of TIMSS, PISA, or PIRLS in their analysis. Articles that used or were inspired by assessment materials or designs from these studies but did not use publicly available data from these studies were excluded. This decision was made because these articles

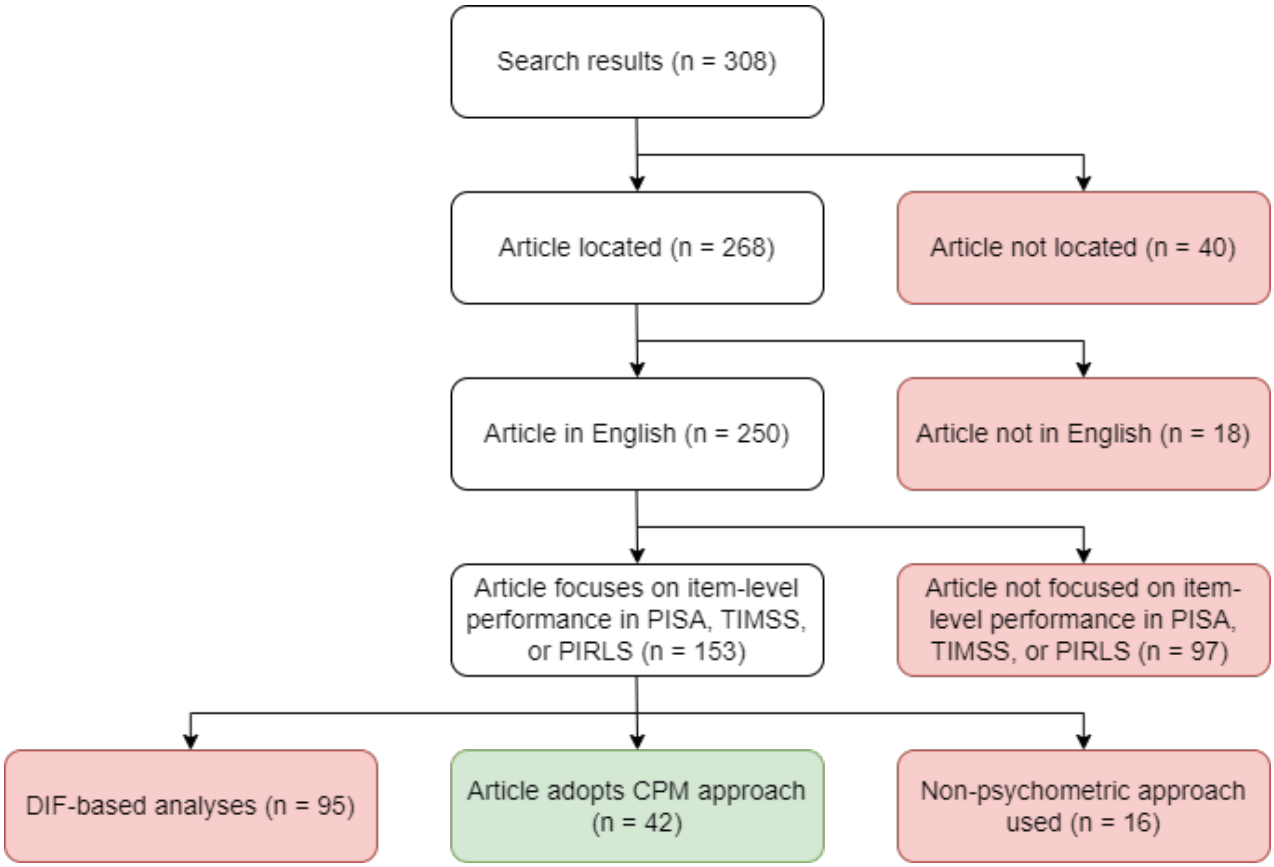
typically included very limited sample sizes or limited subsets of items, and unlike the other articles in the review, would not be easily replicated due to limitations with data access.

Additionally, there were a large number of articles located that focused on differential item functioning analyses – this was to be expected given that three of the search terms in List B focused on these types of models. However, these were omitted from the review at a final level of exclusion. There was certainly a rationale for having kept these studies within the review, given that some (though not all) methods of DIF detection calculate and use IRT-based parameters. However, while the findings of DIF analyses can help to highlight potentially problematic items in a test, DIF-based analyses only detect evidence of potential bias - they do not provide explanatory evidence as to what properties of the test-item give rise to such bias. Any explanations for why DIF was found must therefore be made post-hoc, and these can risk overlooking underlying differences in the cognitive profiles of examinees and/or the properties of the test-items, if not supported by more diagnostic statistical methods that account for these (Karami & Nodoushan, 2011). It was my view therefore that these papers did not directly fall under the interest of the research sub-question for this paper.

Nonetheless, DIF-related terms were included in List B for a few reasons; firstly, to allow me to assess the extent to which DIF-related research has been published in comparison to research utilising other CPM models, and secondly, because it was possible that including these search terms would also assist in capturing other relevant research that could be included in the review. Additionally, excluding DIF-related articles at a final level of exclusion allows for a comparative count of how many articles have employed DIF approaches relative to those using CPM approaches. Articles that focused on DIF but that included follow-up analyses using a CPM approach were however included in the review.

A flowchart setting out the process of article exclusion is provided here to show the process of locating and filtering search results (taken from paper 1) is presented in Figure 5. It should be noted that many of the unlocated articles had non-English titles and/or abstracts, and would likely therefore have been excluded at the next stage.

Figure 5: Flowchart of criteria for inclusion in the thematic review



3.1.2 Article coding and synthesis

There were two main stages to the analysis of these articles. In the first stage, basic information about each article was entered into a spreadsheet. This covered basic categorical information such as the ILSA studies, cycles and subjects of focus, information about the national samples included in their analysis, and which CPMs were used. A summary of these is provided in Table 4.

In the second stage of analysis, articles were divided into groups based on the subject focus and/or model selection. For example, there were six articles that used the rule-space method (RSM; Tatsuoka et al., 2004) to analyse TIMSS data, and these were grouped together. Research sub-question 1 looked to identify the main themes and foci of this body of literature, so I decided that it would be easier to synthesise findings by reading the located articles in groups based on the CPMs employed and the subject areas that were targeted. The discussion of the CPMs employed

by previous studies and how they have been applied to ILSAs in paper 1 was broadly shaped around these grouped readings.

Table 4: Summary of article features of interest

Article feature	Summary
Document type	The type of article (e.g. journal article, book chapter, thesis etc.)
ILSA study, cycle, and grade	All of the ILSA studies and their cycles included in the analysis, with the student grade population assessed (e.g. PISA 2000 G9, TIMSS 2007 G8, PIRLS 2006 G4)
Subject focus	All of the subject domains analysed at the item-level (at least one of mathematics, science, reading)
Sample size	A list of all sample sizes included in the analysis (e.g. USA – 4,411; Taiwan – 2,874)
CPMs	A list of all cognitive psychometric models (CPMs) used to conduct analyses (e.g. DINA, LLTM)
Attributes	A list of the attributes / factors / variables used in the CPM analyses, and if adopted from a given source, list what that source was (e.g. 15 attributes representing the 15 subtopic areas of the TIMSS 2007 mathematics assessment framework).

3.2 Papers 2 and 3

Paper 2 focuses on addressing **research sub-question 2**: ***Does the TIMSS mathematics framework possess sufficient predictive power to explain the item-level factors contributing to item difficulty, and if not, how could the framework be improved upon?***

Paper 3 meanwhile predominantly addresses **research sub-question 3**: ***Are the item-level data available from TIMSS suitable for cognitive diagnostic modelling of student performance, and if not, what are the threats to the validity of these analyses?***

Despite focusing on different research questions and ultimately utilising different statistical models, these two papers share the same dataset, have overlapping samples (albeit more limited for paper 3) and rely on the same underlying attribute frameworks used for Q-matrix construction.

Additionally, the statistical models employed in papers 2 and 3 rely on similar assumptions about the structure of the data and the nature of the TIMSS assessment. It is therefore sensible to group some of their discussion here.

All data cleaning and analyses for paper 2 and paper 3 was conducted in R. The eRm package (extended Rasch modelling - Mair et al.; 2024) and GDINA package (Ma & de la Torre, 2020b) were the main tools for analyses in paper 2 and paper 3 respectively.

3.2.1 Datasets

Data for papers 2 and 3 were taken from the 2011 cycle of TIMSS. These data are publicly available from the TIMSS & PIRLS website (https://timssandpirls.bc.edu/). The analyses in both of these papers focused on 85 test-items included in the fourth-grade mathematics assessment that were distributed amongst seven different booklets of test-items.

Paper 2 uses the responses to these items from approximately 20,000 students from six English-speaking countries[3] who participated in TIMSS 2011. The rationale for including data from these six countries was to build a large sample of responses at the item-level. It was important to ensure that all students had completed *exactly* the same items, hence focusing on students who sat the test-items in English rather than countries which administer the test in other languages. A small number of students in Ireland and Northern Ireland who had completed the assessment in Irish were therefore excluded from these analyses. There were other countries that also administered the assessment in English to some or all examinees in their national samples. However, these countries typically had very different levels of average performance to the six included countries (overall and at the sub-domain level, as reported by the IEA; Mullis et al., 2012), and the majority of these were countries where English is a common language of instruction, but not necessarily the common spoken language. Given that paper 2 included analyses relating to readability of TIMSS mathematics items, I decided it would be sensible to

[3] Australia, England, Ireland, Northern Ireland, New Zealand, and the United States, total n = 20,319. However, each individual test-item only has response data from around 25% of the students due to the TIMSS 2011 study design.

exclude all but ‘Anglophone’ countries for whom English was predominantly both a first language as well as the language of testing.

Analyses in paper 3 focus specifically on Australia’s cohort of students ($n = 3{,}497$) who were also included in the analyses for paper 2. The rationale for focusing solely on Australian students in paper 3 was primarily due to a desire to create a more homogenous cohort for drawing ‘diagnostic’ conclusions. This still would have been possible with the wider cohort and would have of course improved item-level response counts, which were a little lower in paper 3 than might have been desired. However, the DINA model is also computationally demanding, and this also was a factor in my selection of a single country’s data for this paper. Australia was selected over the other national cohorts used in paper 2 because it represented a more ‘average’ country in terms of TIMSS 2011 mathematics performance (e.g. compared to England, Northern Ireland and the United States), and also had a larger cohort size than the other more ‘average performing’ countries in this group (e.g. Ireland and New Zealand).

3.2.2 TIMSS items and data structure

One of the main findings I drew from paper 1 (Chapter 4) was that TIMSS has drawn considerably more interest from researchers conducting diagnostic modelling than PISA or PIRLS. This is perhaps surprising given that PISA results draw considerably more interest from educational researchers and policymakers (Hernández-Torrano & Courtney, 2021; Hopfenbeck et al., 2018; Teig et al., 2022; Sjøberg & Jenkins, 2020). However, as discussed in section 2.1, the TIMSS mathematics framework is publicly mapped on to items in greater depth than in PISA. Perhaps most importantly, the IEA has historically released a substantially higher proportion of the items used in TIMSS than the OECD has released for PISA; though this has declined in more recent cycles of TIMSS. The analysis of Q-matrix D (more information in section 3.3.2.2) is dependent upon access to these individual items.

Having selected TIMSS as the study of the choice, the selection of the 2011 cycle of the study may initially appear odd. Data are available from the more recent 2015, 2019 and 2023 cycles of the study. Additionally, TIMSS has now moved away from paper-based assessment to a digital

format. More recent data would be more reflective of current standards and the state of mathematics education and assessment in the six chosen countries. However, the 2011 cycle was chosen primarily because the analyses for paper 2 and paper 3 required access to the items as they were presented to examinees. While the IEA release performance data for all individual test-items used in a cycle, as well as information on what content and cognitive domains they targeted, they do not publicly release every test-item. This is to protect the integrity of the test-items that they intend to use in future TIMSS assessments, as shared items are required for anchoring the data in newer cycles of TIMSS to those of the past and equating performance on the same scale. The vast majority of items from TIMSS 2011 have been released, with 11 of the 14 blocks compared to just six from 2015, and a smaller subset available on request from 2019. Consequently, using data from TIMSS 2011 rather than from a more recent cycle allowed for a greater bank of items and a wider pool of students to be included in the analyses.

In TIMSS 2011, not all students were administered the same mathematics test-items. Instead, each item was assigned to one of 14 blocks (numbered M01 to M14), with each block containing around 10-14 items. Students who participated in TIMSS 2011 were given one of 14 booklets which contained two of these consecutively numbered blocks (e.g. M01 and M02, M07 and M08, M14 and M01 etc.). The reason the IEA adopt this approach is to ensure that the assessment as a whole targets the entire breadth of the TIMSS mathematics assessment framework in sufficient depth, while minimising the demands placed on individual students. As discussed in section 2.1.1, the goal of TIMSS is not to provide accurate measurement of performance at the student level, but rather is optimised to provide accurate measurement at the national level, and therefore this approach is appropriate. This does however introduce additional complexities into data analysis and the IEA's estimation of students' performance. It also introduces complexities into the analyses I have conducted in paper 2 and paper 3.

In both paper 2 and paper 3, it is not possible to include data from all students and item blocks. This is because three of the item blocks used in TIMSS 2011 (M08, M12, and M14) have not been publicly released. While basic information on these items has been released (e.g. content and cognitive domains, estimated difficulty parameters), it is not possible to conduct a detailed

analysis of item content to construct suitable Q-matrices. As such, the analyses in papers 2 and 3 use the 85 test-items that are included in blocks M01 to M07, the largest contiguous chain of released item blocks. My analyses also only include data from the students who were given a booklet that included the test-items from at least one of these blocks (booklets 1-7 and booklet 14), and who had valid response data to at least two items – i.e. students who only attempted one test-item or fewer were excluded.

TIMSS 2011 consisted of a mix of multiple-choice and constructed response items. Of the 85 items included in these analyses, 42 items were multiple-choice, while the other 43 required a constructed response. Each multiple-choice item consisted of one correct response and three distractors. Seven of the constructed response items were scored out of two, with a partially correct response being awarded one mark. For the purpose of these analyses, these items were dichotomised, such that only a fully correct response was coded as correct, while partially correct responses were recoded as incorrect.

In addition to the issues associated with retrofitting discussed in Chapter 1, there are also some practical limitations to the use of TIMSS data. The booklet structure used in TIMSS might allow for a greater coverage of items than would otherwise be feasible, but it does produce a large amount of missing data – on average, items only have response data from around a seventh of the actual TIMSS 2011 cohort, and around a quarter of the cohort of students used in my analyses. This affects my analyses in both paper 2 and 3. The data are systematically missing and accounted for in the TIMSS assessment design, but does introduce some complexities into the interpretation of results, as well as in the calculation of some item-level parameters that would not necessarily be introduced with an original test, or with an assessment administered and scored more traditionally (e.g. data from key stage 2 mathematics assessments in England). This is a key limitation for my analyses, particularly in paper 3 which focuses on what we can learn about students from their pool of test-item responses.

3.2.3 Cognitive psychometric models

The analyses presented in paper 2 and paper 3 focus on two main models heavily rooted in IRT traditions, and one model from a cognitive-diagnostic background. These are the one-parameter logistic model (1PL) and the linear logistic test-model (LLTM) in paper 2, and the deterministic-inputs, noisy 'and' gate (DINA) model in paper 3. These have been previously discussed in sections 2.2.4 and 2.2.5 respectively. However, it is important to reiterate that, despite being presented separately in different papers, their findings should be seen as complementary. In particular, the item-explanatory analysis presented in paper 2 using the 1PL and LLTM seeks to help provide validity to the person-diagnostic DINA analyses presented in paper 3.

3.2.4 Q-matrices

One of the key components of many CPM analyses, including both LLTM and DINA analyses, is the Q-matrix, as defined by Tatsuoka (1983). This has been described previously in sections 1.2 and 2.2.2, with an example Q-matrix presented in the former.

Both paper 2 and 3 conduct analyses that are heavily reliant on these Q-matrices. Four Q-matrices are explored in paper 2, while two are used in paper 3. In paper 2, the rationale was to try and identify a Q-matrix that shows strong explanatory fit to the estimates of item difficulty as calculated by the 1PL. The process of creating these Q-matrices, and the rationale for their selection is outlined further in section 3.3.2.

In paper 3, the rationale was to use the strongest fitting Q-matrix from paper 2 to explain examinee performance in terms of attribute mastery and draw comparisons to a Q-matrix that, while based on the TIMSS mathematics assessment framework, may not have as strong a relationship with item difficulties. It can therefore be seen that one purpose of the analyses in paper 2 is to provide support for the validity of the analysis presented in paper 3, which, as I have suggested in paper 1, has been questionable in previous CDM analyses.

The Q-matrices used in paper 2 and paper 3 are strongly related to one another but not identical. While there would be merits to employing the exact same Q-matrices in both papers, the need to

use different Q-matrices arose in part due to computational demands imposed by the DINA model, but also due to differing goals of the two papers. To avoid confusion of the Q-matrices used in paper 2 and paper 3, I have labelled the four Q-matrices in paper 2 with letters (A – D) and numbered the two Q-matrices used in paper 3 (Q-matrix 1 and Q-matrix 2). Q-matrix A and Q-matrix 1 are highly related, while Q-matrix D and Q-matrix 2 are highly related. All six complete Q-matrices (total) are presented in Appendix C.

3.3 Paper 2 analyses

3.3.1 Item-explanatory modelling

The 1PL model was used in paper 2 to calculate independent estimates of item difficulties. These were then compared with the item difficulty estimates produced by the LLTM under each of the four Q-matrices used in the paper. As discussed in section 2.2.4, the LLTM is a variant of the 1PL that introduces restrictions on the calculation of item difficulties, such that each item's difficulty is the linear sum of the difficulty parameters of the attributes associated with that item.

To estimate item difficulties in TIMSS 2011, the IEA used the 2PL for constructed-response items, the 3PL for multiple-choice items, and the partial credit model for polytomously scored constructed-response items (Bezirhan & von Davier, 2024). This of course contrasts with the 1PL approach I have used in paper 2. Differing from the IEA's method of estimating item-level difficulties is not ideal but meets a practical need for the 1PL and LLTM analyses to mirror each other as much as possible.

As previously discussed, the IEA published their own estimates of the difficulty of test-items included in TIMSS 2011. It would therefore be possible to bypass this step and use the IEA's own item parameter estimates. However, paper 2 conducted analyses on a more limited cohort of students from six countries, and additionally, was calibrated solely around the subset of 85 test-items, rather than other items included in TIMSS 2011, as well as those discontinued from previous TIMSS cycles; the IEA calibrates item parameters of new test-items to those of previous cycles through a complex anchoring methodology utilising trend items, allowing for the equating

of all item difficulties on a single scale across time. The estimates of the difficulties of items included in TIMSS 2011 were therefore also based off the estimates of item difficulties in TIMSS 2007, TIMSS 2003, and so on.

To minimise potential effects of differential item functioning and other forms of bias potentially introduced through the different samples and/or selection of items included in this analysis, all item difficulty estimates were newly calculated under the 1PL, and a correlation analysis conducted to assess how well these corresponded to the IEA's estimates.

It is typical when conducting the 1PL to either standardise item difficulties or examinee abilities by setting the mean of one of these to zero. For the analyses presented in paper 2, the mean item difficulty was set to 0, and comparisons of the relative abilities of examinees were made with reference to this item difficulty mean. The normalisation constant introduced in the LLTM was also calibrated to ensure that the resulting mean item difficulty was also equal to 0.

Model fit and 1PL assumption testing was conducted on the split dataset to ensure that the 1PL held for the data. Comparisons of model fit under the IEA's mixed 2PL and 3PL approach were conducted to evaluate the extent to which the reduction to the 1PL impacts overall model fit and the appropriateness of the 1PL. Additionally, item-level fit analyses were conducted to identify whether any individual items no longer fitted well under the 1PL or more generally for this more limited cohort of students relative to the wider TIMSS 2011 cohort. More discussion on overall model fit and item-fit analysis is presented in section 3.3.3.

A high correlation between the item difficulty parameters produced under the 1PL and the LLTM suggests that the proposed Q-matrix of item attributes adequately explains the driving factors of the 1PL's estimates of item difficulty. Poor correlation, however, suggests either that the Q-matrix does not adequately account for the factors contributing to the difficulty of the test-items, or that the assignment of attributes to items in the Q-matrix has not been conducted accurately (Green & Smith, 1987; Baghaei & Kubinger, 2015). Likelihood ratio tests were used to compare the fit of the 1PL and LLTM.

3.3.2 Q-matrices

3.3.2.1 Q-matrices A-C

Four Q-matrices were used to assess data under the LLTM. The first three of these, Q-matrices A – C, were adapted from the mathematics assessment framework utilised by the IEA in their classification of the test-items used in TIMSS 2011 (Mullis et al., 2009) and/or the TIMSSPIRLS 2011 relationships report (Martin & Mullis, 2013). No new coding of the items was required, as this did not require access to the test-item content. These three Q-matrices are outlined in Table 5. Attributes (j) are labelled by their Q-matrix (e.g. A, B, C) followed by a two-digit number (e.g. j_{B04}).

Q-matrix A was based solely on the major cognitive and content domains used by TIMSS in their reporting of results, while Q-matrix B expanded on these by breaking the content domains down further into topic areas and topic bullets within the TIMSS 2011 assessment framework. Q-matrix C added to this by including item-level data from the TIMSSPIRLS relationships report, which explored the readability of the items included in the fourth-grade TIMSS assessment.

Table 5: Outline of attributes in Q-matrices A-C

Q-matrix	Summary of attributes
A	**Six attributes based on the TIMSS 2011 cognitive process and content domains.** --- **Cognitive attributes:** j_{A01} – Knowing, j_{A02} – Applying, j_{A03} – Reasoning --- **Content attributes:** j_{A04} – Number, j_{A05} – Geometry, j_{A06} – Reasoning
B	**Thirteen attributes based on the TIMSS 2011 cognitive domains, nine attributes derived from the content domains, topic areas and topic bullets, and one for whether the item requires a constructed response** --- **Cognitive attributes:** j_{B01} – Knowing, j_{B02} – Applying, j_{B03} – Reasoning --- **Content attributes:** j_{B04} – Number sense / place value awareness j_{B05} – Computations with whole numbers j_{B06} – Problem-solving with whole numbers j_{B07} – Understanding of equivalent fractions j_{B08} – Computations and problem-solving with fractions and decimals j_{B09} – Patterns and relationships j_{B10} – Points, lines, and angles j_{B11} – 2D and 3D Shapes j_{B12} – Data --- **Response attributes:** j_{B13} – Item requires a constructed response
C	**Fifteen attributes, the same as in Q-matrix B but also incorporating two attributes from the TIMSSPIRLS Relationships Report on reading demands.** --- **Cognitive attributes:** j_{C01} – Knowing, j_{C02} – Applying, j_{C03} – Reasoning --- **Content attributes:** j_{C04} – Number sense / place value awareness j_{C05} – Computations with whole numbers j_{C06} – Problem-solving with whole numbers j_{C07} – Understanding of equivalent fractions j_{C08} – Computations and problem-solving with fractions and decimals j_{C09} – Patterns and relationships j_{C10} – Points, lines, and angles j_{C11} – 2D and 3D Shapes j_{C12} – Data --- **Response attributes:** j_{C13} – Item requires a constructed response --- **Reading attributes:** j_{C14} – Item classified as having high reading demands j_{C15} – Item includes at least one use of mathematical / technical language

Q-matrix A consisted of six attributes. These corresponded to the three cognitive domains of the TIMSS mathematics framework and are labelled j_{A01} to j_{A03}. The other three attributes represent the three content domains of TIMSS 2011; 'number' problems, geometry problems, and data-related problems and are labelled j_{A04} to j_{A06}. Each item belongs to one of the three cognitive domains, and one of the three content domains, and therefore, each item in the Q-matrix was defined as assessing two of the six attributes.

Q-matrix B built on Q-matrix A by providing more detailed information about the content domain. Each item used in TIMSS 2011 was also associated with one topic area associated with its cognitive domain (e.g. 'fractions and decimals' in the number domain). Within each topic area, each item was also classified as belonging to one topic bullet, providing a more specific description of the mathematical content of the item. Q-matrix B also included a response attribute to denote whether the item required the student to construct an original response (j_{B13}).

Table 7 presents each topic bullet in the TIMSS 2011 assessment framework (Mullis et al., 2009) that was assessed by at least one of the 85 items included in the analysis, and how these were assigned to the attributes in Q-matrix B (and Q-matrix C). My rationale for deciding on the level of specificity of each attribute in Q-matrices B and C was aimed primarily at ensuring that each attribute was measured by at least *four* items. This cut-off was chosen because Baker (1993) reported misspecification issues associated with LLTM analyses using sparse Q-matrices where some or all attributes were measured by fewer than four test-items, though to my knowledge, there is not an established rule-of-thumb for this cut-off. Further, Baker noted that Q-matrix misspecification arising from sparse matrices was partially offset by larger sample sizes, which the analyses in paper 2 are supported by, so it is possible that this cut-off value is slightly conservative.

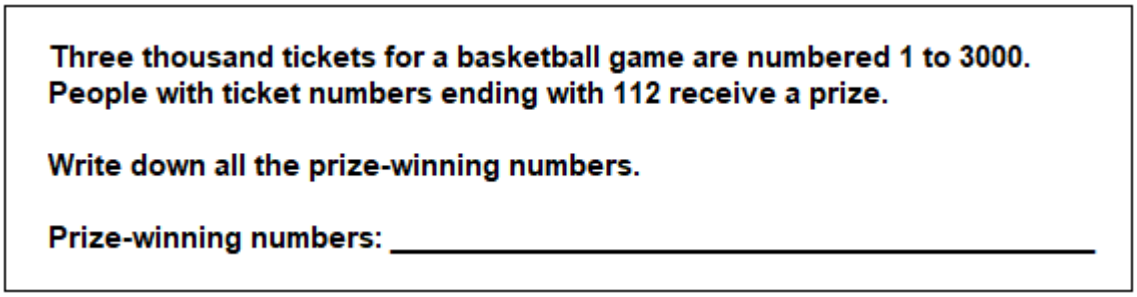
Table 6: Topic bullet mapping to attributes in Q-matrices B and C

Attribute and topic bullets	No. of items
j_{B04}/j_{C04} - **Number sense and place-value awareness**	**8**
Demonstrate knowledge of place value, including recognizing and writing numbers in expanded form and representing whole numbers using words, diagrams, or symbols.	3
Recognise multiples and factors.	1
Find the missing number or operation in a number sentence.	2
Model simple situations involving unknowns with expressions or number sentences.	2
j_{B05}/j_{C05} - **Computations with whole numbers**	**5**
Compute with whole numbers (+, −, ×, ÷) and estimate such computations by approximating the numbers involved.	5
j_{B06}/j_{C06} - **Problem-solving with whole numbers**	**17**
Solve problems, including those set in real-life contexts and those involving measurements, money, and simple proportions.	17
j_{B07}/j_{C07} - **Understanding of equivalent fractions**	**4**
Show understanding of fractions by recognising fractions as parts of unit wholes, parts of a collection, locations on number lines, and by representing fractions using words, numbers, or models.	1
Identify equivalent simple fractions; compare and order simple fractions.	3
j_{B08}/j_{C08} - **Computations and problem-solving with fractions and decimals**	**4**
Add and subtract simple fractions.	1
Add and subtract decimals.	1
Solve problems involving simple fractions and decimals.	2
j_{B09}/j_{C09} - **Patterns and relationships**	**8**
Extend or find missing terms in a well-defined pattern, describe relationships between adjacent terms in a sequence and between the sequence number of the term and the term.	7
Write or select a rule for a relationship given some pairs of whole numbers satisfying the relationship, and generate pairs of whole numbers following a given rule (e.g. multiply the first number by 3 and add 2 to get the second number).	1
j_{B10}/j_{C10} - **Points, lines, and angles**	**9**
Measure and estimate lengths.	1
Compare angles by size, and draw angles (e.g. a right angle, angles larger or smaller than a right angle).	3
Use informal coordinate systems to locate points on a plane.	5
j_{B11}/j_{C11} - **2D and 3D Shapes**	**19**
Identify, classify and compare common geometric figures (e.g. classify or compare by shape, size, or properties).	3
Recall, describe, and use elementary properties of geometric figures, including line and rotational symmetry.	10
Recognise relationships between three-dimensional shapes and their two-dimensional representations.	2
Calculate areas and perimeters of squares and rectangles; determine and estimate areas and volumes of geometric figures.	4
j_{B12}/j_{C12} - **Data**	**11**
Read scales and data from tables, pictographs, bar graphs, and pie charts.	3
Compare information from related data sets.	2
Use information from data displays to answer questions that go beyond directly reading the data displayed.	3
Compare and match different representations of the same data.	2

While two of the attributes in Q-matrix B were measured by exactly four test-items, there are other attributes such as j_{B11} (2D and 3D shapes) that were measured by considerably more. In this case, for example, I did not further separate this attribute into different multiple related attributes because some of the individual topic bullets in the TIMSS assessment framework did not have the sufficient coverage to be included in isolation.

As an example of how Q-matrix B related to individual test-items, Figure 6 shows an example of one of the 85 test-items included in the analyses in paper 2. It was categorised as testing the reasoning cognitive domain and the number content domain. Within the number content domain, it belonged to the 'whole numbers' topic area. It was also classified as requiring the student to 'solve problems, including those set in real-life contexts and those involving measurements, money, and simple proportions'. Under Q-matrix B, it was therefore assigned attributes: j_{B03} (reasoning) and j_{B06} (problem-solving with whole numbers). This item also required students to write down a list of numbers, rather than selecting one of four multiple-choice options, and was therefore also assigned attribute j_{B13} (constructed-response).

Figure 6: Example item 49 (M031016)



Source: TIMSS 2011 Assessment. Copyright © 2013 International Association for the Evaluation of Educational Achievement (IEA). Publisher: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Q-matrix C introduced two additional attributes taken from the TIMSSPIRLS 2011 relationships report (Martin & Mullis, 2013). One of the analyses presented in that report focused on reading-related contributions to item difficulty in TIMSS mathematics and science, and explored relationships between reading performance (as estimated by performance in PIRLS) and item-

level performance in TIMSS. Reading-related attributes in their analysis were either extracted directly from the items themselves (e.g. word counts, number of symbolic displays) or judged 'holistically' by a team of 10 context experts (e.g. overall reading demands). Item-level information on this coding is available from the TIMSSPIRLS relationships report page of the TIMSSPIRLS website (https://timssandpirls.bc.edu/timsspirls2011/international-database.html). The two attributes included in Q-matrix C were selected because they were deemed the most suitable to be applied to a Q-matrix.

The first attribute introduced in Q-matrix C (j_{C14}) was whether an item was judged by the team of content experts to have high reading demands or not. Of the 85 items included in this analysis, 30 items were judged to have high reading demands, 25 were judged to have medium demands, and 30 were judged to have low reading demands. Under Q-matrix C, a binary distinction was made between items with high reading demands and non-high (i.e. medium or low) demands.

The second attribute introduced in Q-matrix C (j_{C15}) was whether the item included at least one example of technical or mathematical language (otherwise referred to in the report as 'specialised vocabulary'. This applied to 47 of the 85 items included in this analysis. The IEA reported on the total count of what was deemed 'specialised vocabulary' for each item, but not *what* those terms were.

While instances of technical language formed a part of the judgement as to whether an item had high reading demands, it should be noted that only 13 of the 30 items with high reading demands included an example of technical language. However, there was a strong association between reading demands and whether the item covered the data content domain; of the 11 data items, nine were judged to have high reading demands, with just one medium demands and one low demands item. The TIMSS 2011 relationships report (Martin & Mullis, 2013) notes that this is primarily due to the majority of these items including dense visual displays with many elements (e.g. different graphs and tables).

3.3.2.2 Q-matrix D

Following from these three Q-matrices, model fit was explored through an additional Q-matrix that adds greater specificity regarding the cognitive processes and mathematical skills elicited by the items. This required a process of item coding that goes beyond how the IEA classified the items used in TIMSS 2011. This process involved the selection of an alternative framework for selecting relevant attributes for consideration, an initial phase of Q-matrix construction, followed by revision of the Q-matrix to ensure sufficient coverage of attributes to items.

Q-matrix D was adapted from the Q-matrix first used by Tatsuoka et al. (2004) in their exploration of attribute mastery of examinees in the eighth-grade assessment of TIMSS-R 1999. This Q-matrix was chosen because it influenced much of the early CDM research using data from international assessments (e.g. Birenbaum et al., 2005; Chen et al., 2011), and unlike more recent CDM research that has typically focused on using the TIMSS framework only, was instead based on an analysis by of student's written responses to test-items conducted in collaboration with a pair of secondary school teachers. Their Q-matrix was split into 23 attributes clustered into three groups; content attributes, process attributes, and skill attributes. However, given the differing ages of the cohorts, as well as changes to the IEA's framework of item development over time, some of the aspects in their Q-matrix were not well tailored to the items in the fourth-grade assessment in 2011. Additionally, on inspection of a first draft of the constructed Q-matrix, some attributes were assigned to less than three items, and were subsequently removed or reintegrated with a similar, more well represented attribute.

The final list of attributes derived from Tatsuoka et al.'s Q-matrix is reported in Table 7. Each item included in this analysis was coded to belong to one of the five content attributes, and most belonged to a single process attribute, though some items were classified as belonging to more. Items could be classified as testing any number of the skill attributes (range 0-5).

Table 7: Outline of attributes in Q-matrix D

Attribute	Description
Content	
j_{D01}	Mathematical concepts and operations involving whole numbers
j_{D02}	Mathematical concepts and operations involving fractions and decimals
j_{D03}	Geometry content
j_{D04}	Data-related content
j_{D05}	Measuring or estimating in real-world contexts: (e.g. lengths, time, temperature etc.)
Process	
j_{D06}	Translate/formulate equations and expressions
j_{D07}	Perform computations in arithmetic / geometry
j_{D08}	Make judgements or evaluate mathematical correctness
j_{D09}	Logical reasoning in written contexts
j_{D10}	Logical reasoning in visual contexts
Skill	
j_{D11}	Question requires a demonstration of number sense / place-value awareness
j_{D12}	Question requires the student to identify and understand numerical patterns and sequences.
j_{D13}	Question requires the student to identify and understand proportional relationships.
j_{D14}	Question requires the student to compare the properties of different entities in the item.
j_{D15}	Question requires interpretation of information presented in figures, tables, or charts/graphs
j_{D16}	Question requires the student to check/evaluate from a set of options (not just multiple-choice, requires active selection of most appropriate option etc.)
j_{D17}	Question requires an open-ended response
j_{D18}	Question requires verbal interpretation
j_{D19}	Question is embedded in a novel or non-routine context (usually in a real-world context)

Figure 6 (shown in section 3.3.2.1) displays an example item from TIMSS 2011. Under Q-matrix D, this item was coded as targeting the following attributes:

- j_{D01} – the item deals with whole numbers.
- j_{D09} – tests higher-level verbal reasoning.
- j_{D11} – related to place-value understanding.
- j_{D17} – open-ended response.
- j_{D18} – requires verbal interpretation.
- j_{D19} – is embedded in a real-world context.

3.3.2.3 Reliability checks

Misspecification of the Q-matrix is one of the major threats to the validity of CPM analyses (Gu & Xu, 2021; Tatsuoka, 2004; Rupp, 2007; Williamson, 2023). The attributes in the Q-matrix may itself be poorly aligned to test-item content (i.e., they are not the 'correct' attributes), or an appropriate set of attributes may not appear to fit the data well if there are inaccuracies in the assignment of these attributes to test-items.

In Q-matrix D, the assignment of attributes to items is made solely by reviewer judgement. Given that the items used in TIMSS were not designed around these Q-matrices, there is, therefore, a high level of potential ambiguity in how attributes relate to items. Ideally, this assignment of Q-matrix attributes is conducted collaboratively by a team of content experts (Baghaei & Kubinger, 2015). This was not practical for the purpose of my thesis, and I instead enlisted the help of two 'coders'. These coders had doctoral backgrounds in mathematics and statistics but did not have previous experience directly with TIMSS or mathematics test development and analysis.

Checks were conducted on a sample of 30 items, just over a third of the items used in the main analysis. These were conducted by the two coders independently. First, each coder was given a common group of 10 items and asked to code the items with respect to the attributes in Q-matrix D. This was to establish a baseline reliability of coding and ensure that all three coders had interpreted the Q-matrix framework in the same way. Where coders had disagreements either with the original attribute coding, or with each other, the three of us met to discuss the codes we had all assigned and clear up ambiguities in the interpretation of the coding guidance outlined in Appendix D, and ensure that a consensus agreement on codes had been reached. This protocol was followed in the next step.

After each of the coders had completed their first set of coding, they were each given a second sample of 10 items, this time unique to each coder, to ensure a broader coverage of reliability checks. Across the 20 items included in this stage, there were just six disagreements with attributes compared to those I had originally. These disagreements were once again discussed as a group of three. After discussing these disagreements, only one of these resulted in a change

to the Q-matrix, with the original attribute codes being retained in the five other cases. The application of the Q-matrix was generally regarded as being a simple process by these external coders, and the low level of disagreement after the original collaboration suggests that the original attribute coding was at least relatively accurate. Nonetheless, it would have been beneficial to have had coding reliability checks for a greater pool of items, more checks per item, and coding checks by people with prior experience in mathematics assessment. The instructions document sent to reliability coders is presented in Appendix B.

3.3.3 Assessments of model fit

3.3.3.1 Overall model fit

As discussed in section 3.3.1, there was a need to consider the impact of using reduced parameter models compared to the 2PL / 3PL / partial credit models used by the IEA, as well as the focus on a more limited cohort of students. This was done by assessing the correlation of 1PL estimates on the sub-cohort with item-level difficulty parameters released by the IEA. This is presented in paper 2 and showed a strong correlation between the difficulty estimates, indicating that the overall 'loss' from these data reductions was small.

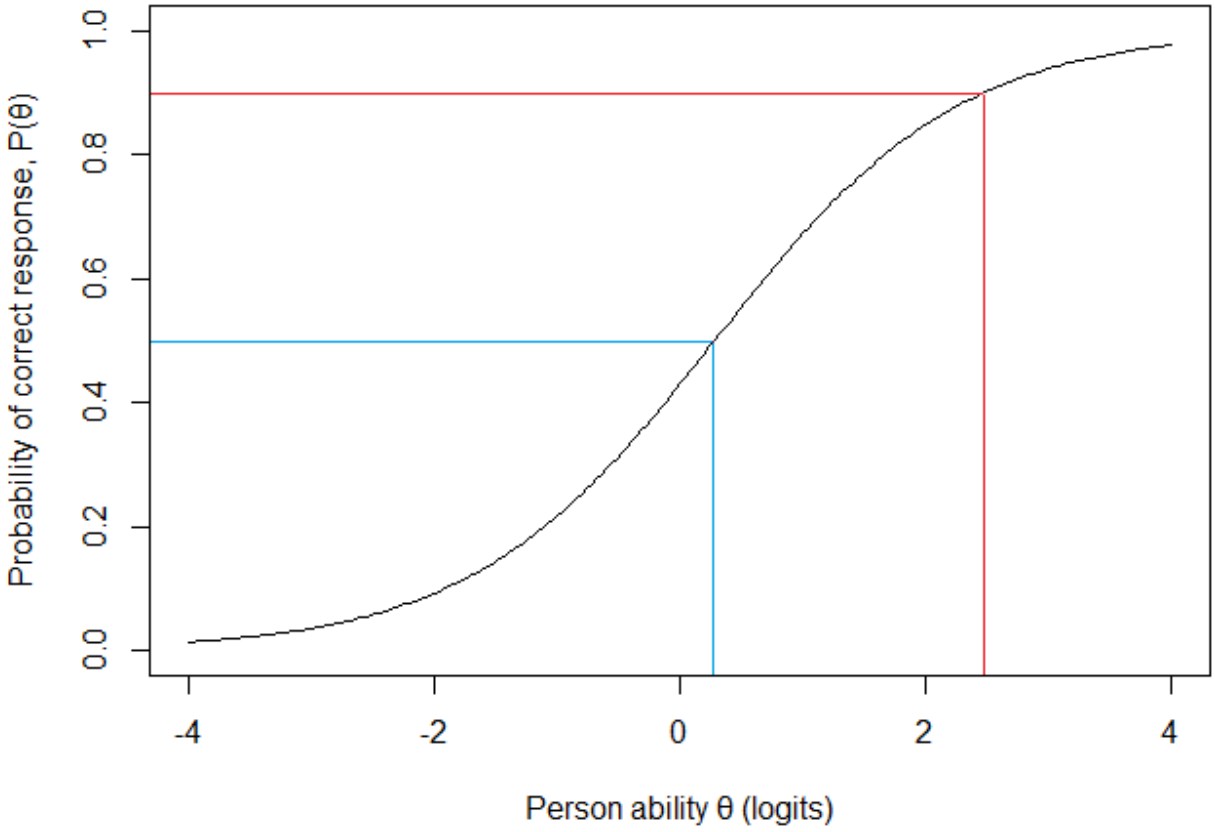
3.3.3.2 Item-level fit

Fit was also explored at the item-level. If there were items that fit the data particularly poorly, these could have been removed from the analysis altogether. The items used in TIMSS 2011 had already undergone such procedures, and given the overall strong correlation with IEA estimates, I anticipated that the number of poorly fitting items would be small. Nonetheless, this warranted a check given the more limited sample of test-items and examinees for parameter estimation, and the dichotomisation of previously polytomously scored items. Two main approaches to assessing item-level fit were adopted.

One method of exploring item fit was through inspection of item-characteristic curves, which plot the probability of success on a given item as a function of examinees' estimated abilities and the estimated difficulty of the item. An example item-characteristic curve from a 1PL analysis used in paper 2 is presented in Figure 7, showing an item with a clear sigmoidal curve. The x-axis shows

the scale of examinee (person) abilities (θ), while the y-axis shows the predicted probability of an examinee of a given ability level providing a correct response to the item, $P(\theta)$. The curve suggests that an examinee with an estimated ability of 0.3 logits would have around a 50% chance of providing the correct response, rising to around 90% for an examinee with a θ ability estimate of just over 2 logits. These positions are represented graphically by the blue and red lines respectively.

Figure 7: Item characteristic curve for item 47 (M051077), 1PL model



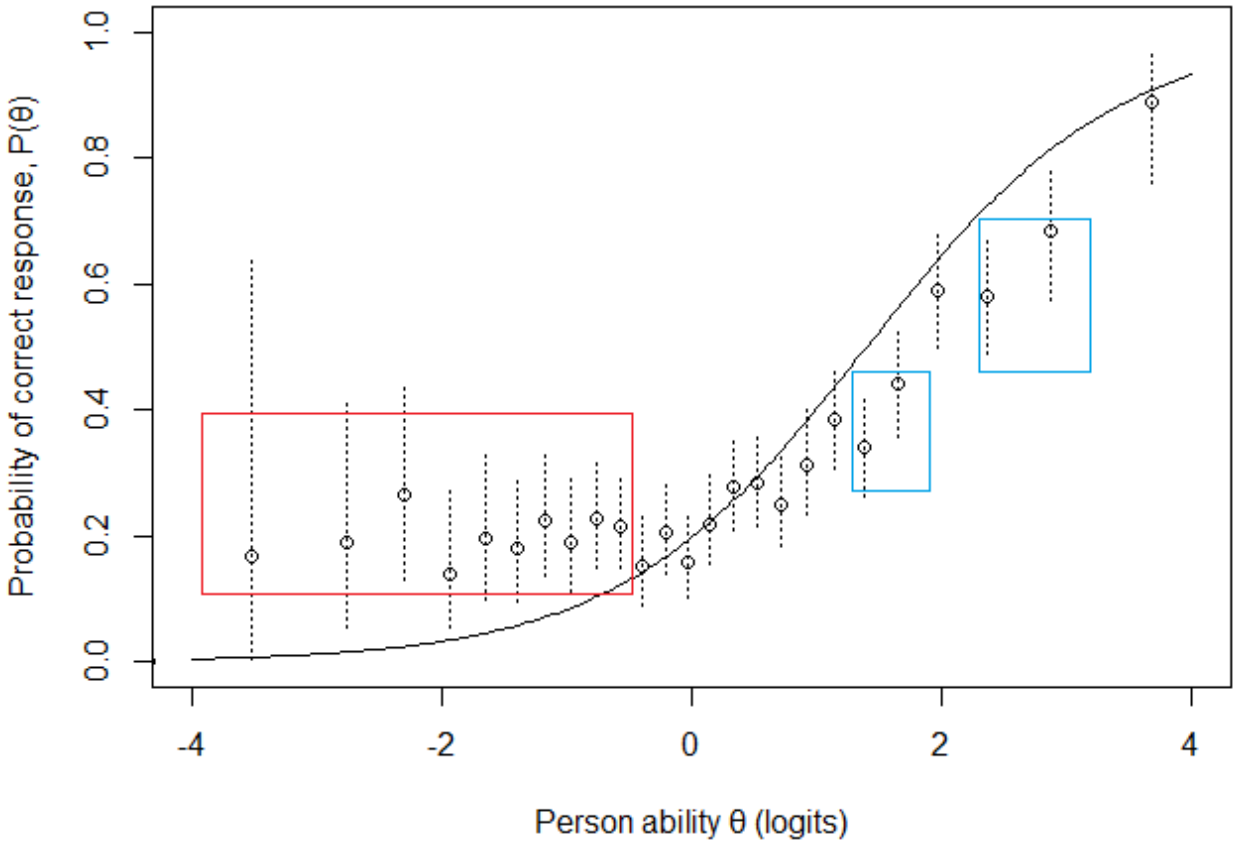
Easier or more difficult items that still fit the data well will have similar shaped curves, but will be translated left or right respectively. Additionally, the slope of the curve may vary when using logistic models such as the 2PL or 3PL that calculate item-level discrimination and pseudo-guessing parameters. Items with curves that deviate substantially from the ideal sigmoidal curve are likely to warrant removal. For example, if an item's difficulty is poorly targeted to the ability of examinees (i.e. almost all or almost no examinees provide the correct response), then the item characteristic curve might appear like a straight line, rather than showing a curve.

Secondly, outfit statistics were calculated for each item. These statistics test the assumption of monotonicity (that an increase in examinee ability in the latent trait of the assessment, i.e. mathematics, is associated with an increase in the probability of success on a given item). An outfit residual of 1 indicates that an item fits the 1PL perfectly. Values less than this suggest that the item may be overfitted to the data. Conversely, items with outfit residuals greater than 1 suggest underfitting to the data, and are generally seen as more problematic. One common threshold is to investigate items with outfit residuals greater than 1.3 (Müller, 2020), though depending on circumstances and sample sizes, other cut-off values may be used.

Outfit residuals were estimated separately for each of the included booklets (M01 to M07, and M14). The reason for conducting this analysis separately was that these analyses cannot deal with the high levels of missing data in the combined dataset. Thus, two outfit residuals were calculated for each item, corresponding to the samples for each of the two booklets that item appeared in. If either residual was greater than 1.3, the item was investigated further through detailed exploration of item characteristic curves overlaid with item probabilities for groups of students based on their raw scores in each booklet. This is useful for identifying groups of students who deviate from the curve; for example, if an otherwise low-performing group overperformed on a given item relative to the curve.

An example of an item demonstrating potential misfit is shown in Figure 8. This is a relatively difficult item (around 1.5 logits) with a high outfit residual (outfit residual ¿1.71). In this example, examinees with a low number of overall correct responses had a higher-than-expected probability of producing the correct response to the item. These examinees are highlighted in the red box. Conversely, examinees with relatively high overall scores (highlighted in the blue boxes) had a lower-than-expected probability of producing the correct answer. This deviance from the 1PL curve at both the lower and upper end of the probability distribution is why the overall outfit residual for the item is high. A similar pattern was observed for the group of students who answered this item in the other booklet, therefore suggesting this could be a consistent pattern for this item. Outfit statistics are briefly discussed in paper 2, and presented in greater detail in Appendix D.

Figure 8: Example item characteristic curve of an item with high outfit residuals



One item from this analysis was flagged as potentially problematic, with two separate outfit residuals greater than 1.5. This could have warranted the removal of this item from further analyses. On further inspection of the item and the relation to raw item scores, I decided to retain the item for the LLTM analyses; I believe that this decision is justified given that this item *was* retained by the IEA for scoring in TIMSS 2011, and because the 1.3 cutoff value may be seen as somewhat conservative (e.g. Linacre, 2002; Bond & Fox, 2015).

Previous research has suggested that for assessments where item difficulties are estimated under the 1PL or Rasch models, item misfit is more likely to arise in assessments that include a mixture of both multiple-choice and constructed-response items (e.g. Andrich, Marais & Humphry, 2016; Andrich & Marais, 2018). This is because guessing behaviours are more likely to bias item difficulty estimates of multiple-choice items, but this cannot easily be corrected for linearly because of the inclusion of constructed-response items. TIMSS uses a mixture of the 2PL and 3PL for the estimation of item difficulties, the latter of which includes an item-level guessing parameter. In paper 2, I used the 1PL instead, for its greater comparability with the LLTM, and I

therefore did not control for guessing behaviour. The overall correlation of item difficulties as estimated under the 1PL and 2PL/3PL approach was strong ($r = .87$), helping to justify the use of the 1PL in the paper. However, it should be noted that an approach which better controlled for potential guessing behaviour could have reduced item-level misfit further. One potential approach, as demonstrated by Andrich and Marais (2018), could have been to post-hoc recode the responses of students who gave correct answers to multiple-choice items that were estimated to be very difficult relative to their ability from correct to missing. The model could then have been rerun with these more anomalous responses excluded. For the 1PL or Rasch models, which do not include a separate item discrimination parameter, this recoding leads to changes in item difficulty estimation for both multiple-choice *and* constructed-response items. Andrich and Marais suggested that this approach can stabilise item difficulty estimation and reduce potential item misfit, particularly in assessments with mixed response types.

3.3.3.3 Relative 1PL and LLTM fit

In paper 2, likelihood ratio tests were conducted to assess the relative fit of the 1PL and LLTM. If the proposed framework of attributes in a Q-matrix used for the LLTM insufficiently accounts for the item-level features contributing to item difficulty, or if there is misattribution of the framework, then the LLTM will show poor fit relative to the 1PL. However, likelihood ratio tests alone can be misleading, particularly when retrofitting a Q-matrix to data from pre-existing assessments (Baghaei & Kubinger, 2015; Fischer & Foreman, 1982), and particularly when sample sizes are large; this is because likelihood ratio tests are quite powerful. As such, it is highly unlikely (both in this analysis and more broadly) that the LLTM will exhibit greater fit to the data than the 1PL.

3.4 Paper 3 analyses

3.4.1 Q-matrices

The goal of paper 3 was to explore issues in the application of the DINA model to TIMSS data. One of the main areas of discussion are issues associated with Q-matrix construction for these kinds of analysis, and how these issues manifest in their reports of students' mathematical masteries. Gu and Xu (2021) provided a recommendation of three features that should be used

to evaluate the suitability of a Q-matrix for cognitive diagnostic modelling. In paper 3, I evaluate two different Q-matrices with respect to these features.

The first of these is Q-matrix 1, which was adapted from Q-matrix A in paper 2. Q-matrix A consisted of just six attributes, which would make it relatively suitable for the DINA model, as the DINA model performs better when there are a lower number of attributes under investigation (e.g. Ravand & Baghaei, 2020). However, one of Gu and Xu's recommended criteria is 'completeness'. Completeness is achieved if there is at least one item for each attribute that measures it in isolation (i.e. an item that measures that attribute, and no other attribute). This is preferred because it helps lead to more stable parameter estimation for that attribute. However, every item in Q-matrix A was coded as having measured exactly two attributes (the associated cognitive domain attribute and the associated content domain attribute). To ensure that at least one of the two Q-matrices used in paper 3 could be argued to be 'complete', I followed the approach recommended by Madison and Bradshaw (2015), who suggested that composite attributes are created that represent the interaction of those attributes. In this case, the composites were the interaction between the content and the cognitive domain measured by each item. This approach had previously been used by George and Robitzsch (2018) in their application of the DINA model to TIMSS data.

Q-matrix 1 therefore consists of nine attributes, with each test-item measuring one attribute. All nine attributes represent the domain interaction of that test-item (e.g. number*reasoning). The counts of the number of items assessing each of these attribute interactions is reported in Table 8.

Table 8: Counts of interactions of content and cognitive domains in Q-matrix 1

	Cognitive Domains			
Content domains	*Knowing*	*Applying*	*Reasoning*	∑
Number	15	19	12	46
Geometry	12	13	3	28
Data	4	4	3	11
∑	31	36	18	85

Q-matrix 2 meanwhile was adapted from Q-matrix D. Q-matrix D consisted of 19 attributes. This posed a number of problems when looking to apply this Q-matrix to the DINA model, both theoretical and practical. As previously mentioned, the DINA model is vulnerable to poor parameter estimation when the number of attributes is large (e.g. Ravand & Baghaei, 2020; Şen & Cohen, 2021). In the context of the DINA model, even Q-matrix 1 is large with its nine attributes. Nineteen is therefore far greater than what would typically used for a cognitive diagnostic modelling analysis. Secondly, even for a relatively simple CDM where only a few parameters are calculated per item, the number of calculations required increases exponentially with each additional attribute. This makes even the simpler CDMs very computationally demanding with large numbers of attributes. For paper 3, I decided to reduce the number of attributes from nineteen in Q-matrix D to twelve for Q-matrix 2. This is still on the high-end for the DINA model (Ravand & Baghaei, 2020; Jang, 2009) but is nonetheless more appropriate than the original nineteen attributes. The actions and rationale for reducing the number of Q-matrix 2 was as follows:

- Combining the ‘number’ and ‘fractions and decimals’ attributes of Q-matrix D into a single attribute, as I found in paper 2 that there was a negligible difference between these two attributes in their contributions to item difficulties.
- Combining the ‘reasoning in written contexts’ and ‘reasoning in visual contests’ attributes of Q-matrix D into a single ‘reasoning’ attribute, as there were a somewhat low number of items that assessed one of these, and the difference in their contributions to item difficulty estimates in paper 2 was small.
- Removing the ‘requires verbal interpretation’ and ‘set in novel/non-routine contexts’ attributes, as these were found to have no significant independent effects on item difficulty estimates in paper 2.
- Removing the ‘translate/formulate equations’ as it was very poorly represented amongst most of the seven booklets (i.e. generally clustered in some booklets and entirely absent in others).
- Removing the ‘use visual information’ and ‘open-ended response’ attributes, as these are more item-focused attributes rather than relating to students’ skills or knowledge states.

None of these transformations or removals required further access to the test-items, and therefore no further reliability checks on item coding were required. The final list of twelve attributes included in Q-matrix 2 are listed in Table 9.

Table 9: Summary of attributes in Q-matrix 2

Attribute	Description	No. of items
	Content attributes	
j_{201}	Mathematical concepts and operations involving whole numbers, fractions or decimals	42
j_{202}	Geometry content	26
j_{203}	Data-related content	10
j_{204}	Measuring or estimating in real-world contexts: (e.g. lengths, time, temperature etc.)	7
	Process attributes	
j_{205}	Perform computations in arithmetic / geometry	28
j_{206}	Make judgements or evaluate mathematical correctness	20
j_{207}	Reason logically in verbal and/or visual contexts	21
	Skill attributes	
j_{208}	Question requires a demonstration of number sense / place-value awareness	12
j_{209}	Question requires the student to identify and understand numerical patterns and sequences.	8
j_{210}	Question requires the student to identify and understand proportional relationships.	7
j_{211}	Question requires the student to compare the properties of different entities in the item.	23
j_{212}	Question requires the student to check/evaluate from a set of options (not just multiple-choice, requires active selection of most appropriate option etc.)	21

The rationale for using the attributes in Q-matrix D for person-diagnostic analysis in paper 3 was that the combination of these attributes explained a greater proportion of the variance in item difficulties than the attributes derived solely from the TIMSS mathematics framework. As such, it was theorised that these attributes may draw valid conclusions about students' cognitive profiles. The attribute reduction from Q-matrix D to Q-matrix 2 therefore ran the risk of weakening this relationship between items and attributes. An assessment of the impact of attribute reduction from Q-matrix D to Q-matrix 2 is presented in Appendix E. The loss in explained variance of item difficulties was noticeable but not severe, decreasing from around 42% in Q-matrix D to 36% in

Q-matrix 2. However, Q-matrix 2 still outperformed Q-matrices A, B and C. Overall this suggests that Q-matrix 2 is still suitable for the analysis presented in paper 3.

All analysis was conducted using the GDINA package (Ma & de la Torre, 2020b), which uses a marginal maximum likelihood method with expectation-maximisation algorithm to estimate item guessing and slipping parameters.

3.4.2 Identification of anomalous results

Following a discussion of challenges with identifiable Q-matrix construction for TIMSS test-items, paper 3 draws attention to anomalous results that highlight extreme poor fit of the DINA model to the underlying data. This is done through two main demonstrations; a comparison of the most common latent class profiles to the distribution of the underlying performance data, and an analysis of individual item-level guessing and slipping parameters. The latter of these is reported in full in Appendix F.

I initially planned to compare model fit indices such as the deviance, AIC (Akaike, 1974) or BIC (Schwarz, 1978) of Q-matrix 1 and Q-matrix 2. However, these model fit indices are not reported in paper 3. I made this decision because I do not believe the results of the analysis presented in paper 3 warrant a direct model comparison. This is because it is obvious from the latent class profiles and item-level parameters that both Q-matrices perform *very* poorly and produce nonsensical results. Drawing a comparison to which of the two is “less bad” does not seem sensible when the argument of the paper is that both sets of results should be dismissed entirely.

4 Paper 1: Cognitive psychometric modelling of item-level performance data in PISA, TIMSS, and PIRLS: a thematic review

In a bid to extract more pedagogically useful information from international assessments such as PISA, TIMSS and PIRLS, a number of researchers have looked to the potential applications of cognitive psychometric models to item-level performance data from these studies. This thematic review identifies and discusses 42 articles that applied diagnostic or explanatory cognitive psychometric models to item-level data from international assessments. Results show that the vast majority of these have targeted data specifically from TIMSS mathematics, as opposed to PISA, PIRLS, or TIMSS science. There has also been a much stronger focus on person-diagnostic models such as the deterministic inputs, noisy 'and' gate (DINA) model, with little attention given to item-explanatory analyses beyond those concerned with differential item functioning. The review calls for a greater focus on analyses that first explore the relationships between items and cognitive attributes before drawing explanatory conclusions about how those attributes reflect students' performance.

4.1 Introduction

Recent decades have witnessed substantial growth in the scale and influence of international large-scale assessments (ILSAs; Addey, 2017; Fischman et al., 2018). Arguably the three most influential and well-known of these ILSAs are the OECD's 'Programme for International Student Assessment' (PISA) study, which assesses 15-year-olds across the world in their ability to apply their knowledge and skills in mathematics, science and reading to real-life problems, and the IEA's 'Trends in International Mathematics and Science Study' (TIMSS) and 'Progress in International Reading Literacy Study' (PIRLS), which are designed to directly assess curriculum knowledge, and are thus aligned to the national curricula of participating countries. Although differing in their subject-matter, testing populations, and assessment designs, these ILSAs share similarly high levels of country participation, draw significant media attention upon the release of

their findings, and are used to inform, shape, and justify changes to national curricula (Gorur, 2017).

However, since their inception, ILSAs have also drawn substantial criticism. These include arguments that despite their claims, these ILSAs tell us relatively little about what students in the participating countries actually know or can do in these subjects, or why students in country *X* appear to outperform students in country *Y* (Wagemaker, 2020). Although each of these ILSAs also score examinee performance across topic areas - for example, TIMSS and PIRLS both calculate scores for examinees on different content domains (e.g. ‘geometry, ‘chemistry’, ‘informational reading’) and across cognitive domains (e.g. ‘knowing’, ‘evaluating’) - these subscales of examinee performance are arguably still too broad to be pedagogically useful to participating countries looking to understand why their country underperforms relative to others. Criticisms have also been directed at the assumption that the test-items used in these assessments measure students' skills invariantly across participating countries or language groups (e.g. Rutkowski & Svetina, 2014; Sandilands et al., 2013).

Despite these criticisms, there is also a recognition of the tremendous potential provided by ILSA data. The datasets from previous cycles of PISA, TIMSS, and PIRLS are freely accessible resources to educational researchers, boasting large, nationally representative samples, and include item-level performance data for each participating student alongside a wide array of matched questionnaire data about the student's beliefs and attitudes, as well as information on their home and schooling backgrounds. These ILSA datasets are therefore ideal for addressing many different research questions, and doing so with analyses that can be backed with high statistical power. Such analyses have the potential to greatly expand upon, either by complementing or evaluating, the results reported by the OECD and IEA.

Much of the current psychometric research utilising ILSA data has focused on methods for detecting evidence of potential bias in their test-items. These ‘differential item functioning' (DIF) approaches are used to identify items which may favour, or disadvantage certain groups of examinees on that item that would otherwise be expected to perform similarly. These approaches

have been used on ILSA datasets to, for example, detect evidence of bias in relation to students' socioeconomic background (e.g. Burkes, 2009; Walzebug, 2014), their gender (e.g. Lan, 2014; Casabianca & Lewis, 2018) or the language in which they complete the test (Stubbe, 2011; El Masri & Andrich, 2020). Although the findings of such studies can help to highlight the weaknesses of the test-items that have been used in ILSAs, DIF-based analyses only detect evidence of potential bias - they do not provide explanatory evidence as to what properties of the test-item give rise to such bias. Any explanations for why DIF was found must therefore be made post-hoc, and these can risk overlooking underlying differences in the cognitive profiles of examinees and/or the properties of the test-items, if not supported by more diagnostic statistical methods that account for these. However, there are alternative methods for analysing ILSA data at the item level that can provide information, descriptive or explanatory, that can help supplement such findings and provide more diagnostic information that identifies reasons for differential performance on test-items based on examinees' strengths and weaknesses, or specific item-level features. This more diagnostic research is less prevalent in published works than those concerning DIF or similar approaches. To date, there has been no attempt to synthesise the findings of such research. This thematic review aims to synthesise the current body of research assessing item-level performance in PISA, TIMSS, and PIRLS using advanced diagnostic modelling approaches, herein referred to as ‘cognitive psychometric models' (CPMs).

4.1.1 Cognitive psychometric modelling

Despite the differences in their aims, CPMs focused on modelling either examinee abilities or the psychometric properties of test-items typically share many similarities in their approaches. For example, a key component of most CPMs is the use of a Q-matrix (Tatsuoka, 1983), which defines the relationship between the set of J dichotomous attributes selected for investigation, and the set of I test-items included in the analysis. These attributes might be defined as skills believed to be assessed by test-items, or features of test-items that are predicted to influence item difficulty. The strength of CPMs and their analyses is thus heavily reliant on the selection and accurate specification of the relevant attributes or specific features and structure of each test-item. This can present a challenge in analysing item-level data from assessments such as PISA,

TIMSS, and PIRLS, where the test-items have not necessarily been designed around rigid item attribute frameworks and where there may be many competing demands on item construction beyond those of their subject matter, such as restraints raised by the need to ensure cross-cultural and cross-lingual comparability.

This thematic review will provide an overview of the current body of research conducting CPM analyses on item-level performance data from PISA, TIMSS, and PIRLS, discussing the types of analyses that have been conducted, the overarching conclusions, and current limitations within the existing body of literature. This is with the aim of providing researchers currently or prospectively interested in working with ILSA data (and particularly those looking to evaluate the validity of these assessments) with a summary of the existing research and the future potential for these kinds of model applications.

4.2 Methods

4.2.1 Databases and search terms

The search for relevant articles was conducted in November 2023 and targeted four digital databases: ERIC, PsycINFO, Scopus, and Web of Science. Searches were made on article titles, abstracts and keywords using the terms provided in **Table 10**, with each search including one term from List A, and one term from List B, e.g. “TIMSS" OR ‘‘Trends in International mathematics and science Study" AND ‘‘CPM" OR ‘‘cognitive psychometric model*ing". The search terms in List B were chosen to cover both a wide range of CPMs employed in item difficulty modelling and cognitive diagnostic analyses, as well as other terminology common to the broader scope of difficulty modelling literature. Additionally, while this review does not intend on covering DIF-based analyses and similar variants to this (e.g. differential text or bundle functioning), these search terms were included for two reasons; first, to assess the relative extent to which articles use DIF approaches compared to articles using other psychometric approaches, and second, because it was plausible that articles focusing on DIF-based approaches could also utilise CPMs that would be of interest to this review. In total, these searches yielded 308 unique results.

Table 10: List of search terms

List A search terms (ILSA studies)
One of:
‘‘PISA" OR ‘‘Programme for International Student Assessment"
‘‘TIMSS" OR ‘‘Trends in International Mathematics and Science Study"
‘‘PIRLS" OR ‘‘Progress in International Reading Literacy Study"
List B search terms (models / statistical approaches)
One of:
“item difficulty”
‘‘response demand"
‘‘item demand"
‘‘cognitive demand"
‘‘text demand"
“cognitive complexity”
‘‘text complexity"
‘‘cognitive analysis"
‘‘diagnostic assessment"
‘‘item model*ing"
‘‘response model*ing"
‘‘CDM" OR ‘‘cognitive diagnostic model*ing"
‘‘CPM" OR ‘‘cognitive psychometric model*ing"
‘‘DINA" OR ‘‘DINO" OR ‘‘deterministic input noisy"
‘‘RSM" OR ‘‘rule space method" OR ‘‘rule-space method"
‘‘RUM" OR ‘‘unified model" OR ‘‘reparameteri*ed unified model"
‘‘Fusion model"
‘‘LLTM" OR ‘‘linear logistic test model"
‘‘LPCM" OR ‘‘linear partial credit model"
‘‘automated item generation"
‘‘item characteristics"
‘‘design matrix"
‘‘Q-matrix" OR ‘‘Q matrix"
‘‘evidence cent*red design" OR ‘‘evidence-cent*red design"
‘‘DIF" OR ‘‘differential item functioning"
‘‘DTF" OR ‘‘differential text functioning"
‘‘DBF" OR ‘‘differential bundle functioning"

4.2.2 Inclusion criteria

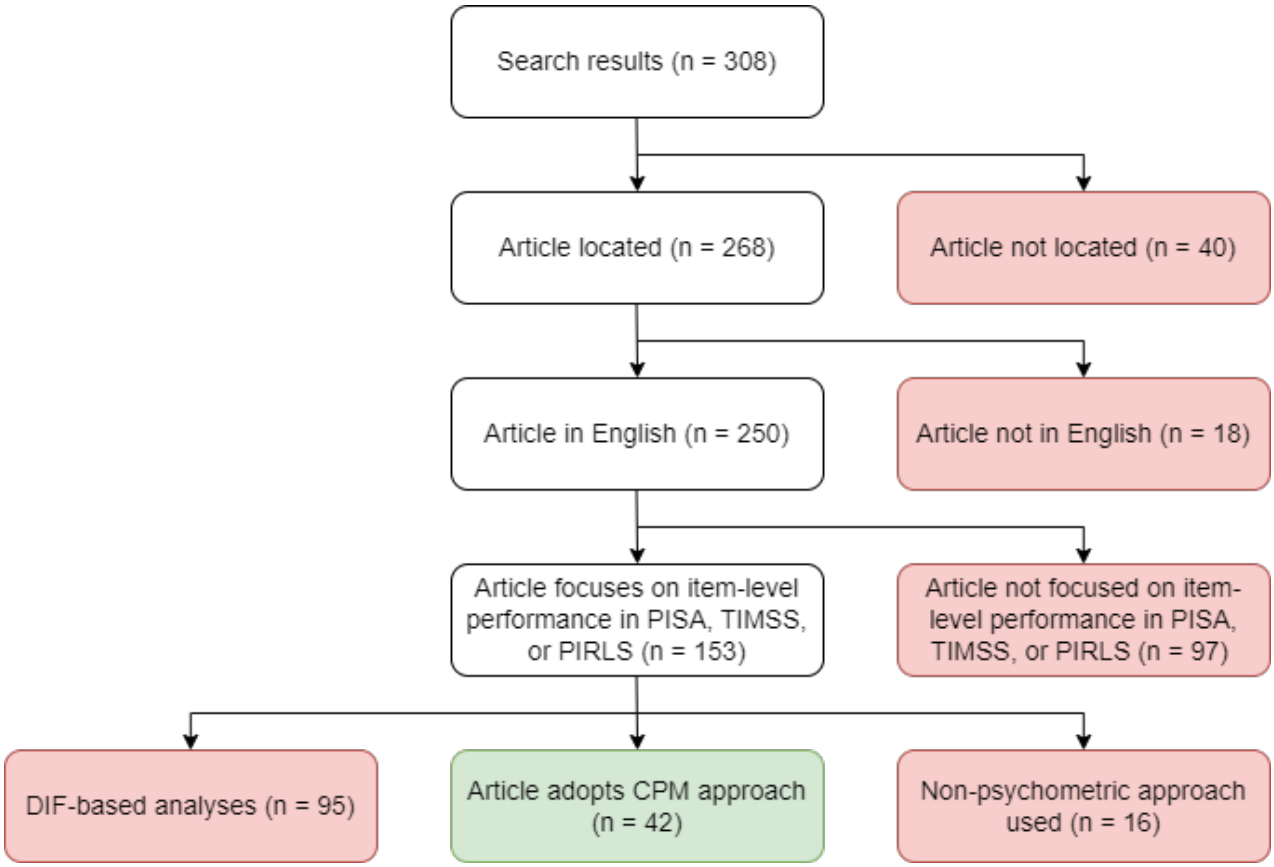
The number of results for each search was recorded, and a running list of articles was kept. Additionally, where available, a copy of each article was downloaded and stored electronically. As expected, a large number of the located articles were only tangentially relevant to this review. In order to filter the articles to identify those of relevance to the review, a set of inclusion criteria

were established. Articles and conference papers were read in full to determine inclusion, but the readings of longer reports, as well as academic dissertations and theses were restricted to targeted chapters selected from contents lists.

For inclusion, a copy of the article needed to be located, be presented in English, and conduct statistical analyses relating to item difficulty on a dataset from a PISA, TIMSS, or PIRLS study. Articles conducting analyses exclusively on other national assessments, those using ILSA assessment materials on alternative student samples, or those using simulated data were not included in the review. Additionally, analyses that did not focus on performance at the item-level in some capacity, such as those that focused exclusively on aggregated plausible values of examinee ability were also not included. Just over half of the articles were omitted after applying these inclusion criteria. Finally, articles were restricted to those that conducted analyses that could be considered ‘cognitive psychometric modelling'; a small number of articles that conducted qualitative analyses, analyses focusing on imputation methods for ILSAs, articles focusing on simple comparisons of mean-level performance for different types of test-items, as well as regression analyses that did not incorporate item difficulty estimates derived from psychometric approaches (e.g. Rasch difficulty estimates) were therefore not included. Finally, articles that only focused on DIF or related forms of differential functioning analyses were excluded separately.

The full process for filtering the articles is outlined in Figure 9, which provides the number of articles excluded at each stage. The separate process for eliminating articles focusing on DIF-based analyses was done to allow for a comparison figure for the number of articles that adopted these approaches to those using CPM approaches. Forty-two articles met all the inclusion criteria. These 42 articles included 29 articles published in academic journals, one conference paper, one report, and the remaining 11 were academic dissertations or doctoral theses.

Figure 9: Flowchart of criteria for inclusion



4.2.3 Analytic approach

All 42 of the relevant articles were coded based on a set of article features of interest, including the ILSA study (and for TIMSS, age cohort) of focus, the types of CPM employed, and where relevant, a summary of what attributes were used to explain properties of items and/or examinees. This set of article features is outlined in Table 11.

Table 11: Summary of article features of interest

Article feature	Summary
Document type	The type of article (e.g. journal article, book chapter, thesis etc.)
ILSA study, cycle, and grade	All of the ILSA studies and their cycles included in the analysis, with the student grade population assessed (e.g. PISA 2000 G9, TIMSS 2007 G8, PIRLS 2006 G4)
Subject focus	All of the subject domains analysed at the item-level (at least one of mathematics, science, reading)
Sample size	A list of all sample sizes included in the analysis (e.g. USA – 4,411; Taiwan – 2,874)
CPMs	A list of all cognitive psychometric models (CPMs) used to conduct analyses (e.g. DINA, linear logistic test model)
Attributes	A list of the attributes / factors / variables used in the CPM analyses, and if adopted from a given source, list what that source was (e.g. 15 attributes representing the 15 subtopic areas of the TIMSS 2007 mathematics assessment framework).

In the second stage of analysis, articles were divided into groups based on their shared characteristics. For example, there were six articles that used the rule-space method (RSM; Tatsuoka et al., 2004) that were grouped together. As the aim of this review was to provide a synthesised discussion of the findings of these studies, it was decided that by reading these articles as a group of related articles, it would be easier to draw wider conclusions about how the research in this space has evolved, and the conclusions that the overarching body of cognitive psychometric modelling research of ILSA data has made.

4.3 Article summaries

Table 12, Table 13 and Table 14 provide summaries of each of the 42 articles located in the search based on the subject (mathematics, science, reading) of investigation. Each table is arranged chronologically by year of publication and provides details on the study dataset used (e.g. data from the fourth-grade assessment in TIMSS 2007), national samples included in the

analysis, the specific CPM(s) used to analyse item-level data, and a brief summary of the construction of a Q-matrix (if applicable).

Of the 42 articles identified in the search, the overwhelming majority, 30, were concerned with mathematics, and of these, all but three focused on TIMSS mathematics, as opposed to PISA mathematics. By contrast, only seven articles focused exclusively on science, (six TIMSS, one PISA), and just four focused exclusively on reading (two PISA, two PIRLS). The remaining article (Harrison et al., 2023) focused on separate analyses for mathematics, science, and reading in PISA 2015. However, no articles were concerned with more than one of the ILSAs, with articles typically focusing on a single cycle of the study. It was however, common for studies to include or compare data from their ILSA study with that of other national datasets.

Table 12: Overview of cognitive psychometric modelling studies of TIMSS/PISA mathematics

Reference	Study, cycle and grade	Countries and sample size	CPM models	Q-matrix attributes / factors
Tatsuoka et al. (2004)	TIMSS-R G8	20 countries (total n = 51,435)	RSM	Five content attributes (e.g. 'fractions and decimals', 'data and statistics'), nine process attributes (e.g. 'rule application in algebra'), nine skill/item-type attributes (e.g. 'solving open-ended items').
Birenbaum et al. (2005)	TIMSS-R G8	USA (n = 4,411) Singapore (n = 2,490) Israel (n = 2,092)	RSM	23 attributes taken from Tatsuoka et al. (2004).
Chen et al. (2006)	TIMSS-R G8	Taiwan (n = 2,874)	RSM	23 attributes taken from Tatsuoka et al. (2004).
Dogan & Tatsuoka (2008)	TIMSS-R G8	Turkey (n = 2,900) USA (n = 4,411)	RSM	23 attributes taken from Tatsuoka et al. (2004).
Chen et al. (2010)	TIMSS-R G8	Taiwan (n = 2,874)	RSM	23 attributes taken from Tatsuoka et al. (2004).
Toker (2010)	TIMSS 2007 G8	Turkey (n = 2,900)	LSDM	20 of the 23 attributes taken from Tatsuoka et al. (2004).
Lee et al. (2011)	TIMSS 2007 G4	USA (n = 564) Massachusetts (n = 127) Minnesota (n = 132)	DINA	15 attributes taken from the 15 topic areas within the three content domains of TIMSS 2007 framework; e.g. 'solve problems involving simple fractions and decimals including their addition and subtraction' and 'read data from tables, pictographs, bar graphs and pie charts'.
Chen (2011)	TIMSS-R G8	Taiwan (n = 1,994)	RSM	23 attributes taken from Tatsuoka et al. (2004).
Toker & Green (2012)	TIMSS 2007 G8	Turkey (n = 4,498)	LSDM	20 of the 23 attributes taken from Tatsuoka et al. (2004).
Chen (2013)	TIMSS 2007	Singapore (n = 1,072)	RDINA HO-DINA RHO-RDINA	15 attributes taken from Lee et al. (2011).
Hou et al. (2013)	TIMSS 2007 G8	USA, Massachusetts and Minnesota (total n = 825)	DINA	Three attributes, representing the three cognitive domains of TIMSS G4 (number, geometry, data).

RSM = Rule-space method; **LSDM** = Least-squares distance method; **DINA** = Deterministic Inputs Noisy "AND" Gate model; **RDINA** = Reparameterized DINA; **HO-DINA** = Higher-order DINA; **RHO-RDINA** = Restricted higher-order RDINA; **DINA-H** = DINA with hierarchy; **DINO-H** = DINO with hierarchy.

Table 11: Overview of cognitive psychometric modelling studies of TIMSS/PISA mathematics (continued)

Reference	Study, cycle and grade	Countries and sample size	CPM models	Q-matrix attributes / factors
Su (2013)	TIMSS 2003 G8	USA (n = 1,497) Basque Country (n = 430) Indiana (n = 384) Ontario (n = 703) Quebec (n = 729)	DINA-H DINO-H	15 attributes modified from the Common Core State Standards (2010) and focusing only on items in the number and algebra content domains of TIMSS; e.g. 'reason about and solve one-variable equations and inequalities' and 'compare two fractions with different numerators and denominators'.
Zhao (2013)	TIMSS 2007 G4	USA (1,131)	DINA A-Hierarchy Fusion Model	15 attributes taken from Lee et al. (2011).
Hansen et al. (2014)	TIMSS 2007 G4	USA (n = 564)	HO-DINA	15 attributes taken from Lee et al. (2011).
Park & Lee (2014)	TIMSS 2007 G4	USA, Massachusetts and Minnesota (total n = 825)	RDINA	Seven attributes condensed from the 15 attribute framework used by Lee et al. (2011); e.g. 'fractions and decimals' and 'reading and interpreting... [data]'.
Ma (2014)	TIMSS 2007 G8	All participating countries (n = 31,589)	LSDM	Seven attributes based on item content and cognitive domains, plus three other attributes relating to cognitive complexity and item-type.
Choi et al. (2015)	TIMSS 2003 G8	USA (n = 740) South Korea (n = 439)	DINA	12 attributes adopted from the NCTM Principles and Standards (2000), split across five content strands; number (three attributes), algebra (three attributes), geometry (three attributes), measurement (2 attributes) and data (one attribute).
Şen & Arican (2016)	TIMSS 2011 G8	South Korea (n =368) Turkey (n = 488)	DINA	13 attributes adopted from the Common Core State Standards Initiative (2010), split across the four content domains (number, algebra, geometry, data and chance); e.g. 'applies and extends previous understandings of arithmetic to algebraic expressions...'
Ma & de la Torre (2016)	TIMSS 2007 G8	USA, Massachusetts and Minnesota (total n = 825)	Sequential-process model	Eight of the 15 attributes used by Lee et al. (2011).

DINA-H = DINA with hierarchy; **DINO-H** = DINO with hierarchy; **A-Hierarchy** = Attribute Hierarchy; **HO-DINA** = Higher-order DINA; **RDINA** = Reparametrized DINA; **LSDM** = Least-squares distance method.

Table 11: Overview of cognitive psychometric modelling studies of TIMSS/PISA mathematics (continued)

Reference	Study, cycle and grade	Countries and sample size	CPM models	Q-matrix attributes / factors
Toker (2016)	TIMSS 2011 G8	Turkey (1,225) USA (1,990) Finland (768) Singapore (1,229)	LCA MRM	n/a
Skaggs et al. (2016)	TIMSS 2003 G8	USA, four different samples (n = 175, 350, 675 and 1,250 per item)	GDM	15 attributes distributed across four different attribute lists (four, six, eight, and ten attribute models. Attributes derived from a variety of USA curriculum sources, including the Common Core State Standards, and the Principles and Standards for School mathematics)
George & Robitzsch (2018)	TIMSS 2011 G4	Germany (n = 3,995)	DINA	Nine attributes representing the nine potential combinations of the three content domains (number, geometry, data) and the three cognitive sub-competencies (knowing, applying, reasoning). Thus, each item was only assigned one attribute.
Yamaguchi & Okada (2018)	TIMSS 2007 G4	7 countries (total n = 5,431)	DINA DINO A-CDM LLM R-RUM	15 attributes taken from Lee et al. (2011).
Zhan, Jiao, & Liao (2018)	PISA 2012 G9	Brazil (n = 480) Germany (n = 411) Shanghai-China (n = 393) USA (n = 402)	JRT-DINA	Seven attributes, covering four content domains of PISA mathematics (change and relationships, quantity, space and shape, uncertainty and data) and three problem contexts (occupational, societal, and scientific).
Zhan, Liao, & Bian (2018)	PISA 2015 G9	58 countries (total n = 8,606)	Joint-testlet DINA JRT-DINA	11 attributes, covering four content domains, four problem contexts (‘personal' added from Zhan, Jiao, & Liao (2018)), and three mathematical processes (including ‘formulating situations mathematically' ‘employing procedures etc.' and ‘evaluating mathematical outcomes'.

LCA = Latent class analysis; **MRM** = Mixed Rasch Model; **GDM** = General Diagnostic Model; **DINA** = Deterministic Inputs Noisy "AND" Gate model; **DINO** = Deterministic Inputs Noisy "OR" Gate model; **A-CDM** = Additive cognitive diagnostic model; **LLM** = Linear logistic model; **R-RUM** = Reduced reparameterized unified model; **JRT-DINA** = Joint response-time DINA.

Table 11: Overview of cognitive psychometric modelling studies of TIMSS/PISA mathematics (continued)

Reference	Study, cycle and grade	Countries and sample size	CPM models	Q-matrix attributes / factors
Ma & de la Torre (2019)	TIMSS 2007 G8	USA (n = 1,328)	G-DINA DINA DINO A-CDM	Seven attributes; ‘whole numbers', ‘fractions, decimals, ratio, proportion and percentages', ‘algebraic expressions and equations', ‘geometric shapes', ‘geometric measurement', ‘data organisation and representation' and ‘data interpretation and chance'.
Terzi & Sen (2019)	TIMSS 2011 G8	20 countries (total n = 8,157)	DINA DINO A-CDM G-DINA	13 attributes adopted from the Common Core State Standards Initiative (2010), split across the four content domains (number, algebra, geometry, data and chance); e.g. ‘Applies and extends previous understandings of arithmetic to algebraic expressions...'
Wang & Qiu (2019)	TIMSS 2003 G8	USA (n = 3,710)	DINA mDINA	Five TIMSS content domains; algebra, data, geometry, measurement, and number.
Ma & de la Torre (2020a)	TIMSS 2011 G8	USA (n = 748)	G-DINA	Seven of the nine attributes taken from Park et al. (2017), which was originally based off the TIMSS 2011 framework and from the NCTM (2000). Seven topic areas split over four content domains.
Delafontaine et al. (2022)	TIMSS 2011 G8	Finland (n = 2,397) USA (n = 5,909) Singapore (n = 3,069) Australia (n = 4,240) Tunisia (n = 2,883)	G-DINA	Nine attributes based on the TIMSS 2011 assessment framework.
Harrison et al. (2023)	PISA 2015 G9 (field trial data)	13 countries (total n = 11,960 in mathematics)	GPCM	Three facets (content, situation/context and process) broken down into their respective levels (four, four and three levels respectively).

G-DINA = Generalized DINA; **DINA** = Deterministic Inputs Noisy "AND" Gate model; **DINO** = Deterministic Inputs Noisy "OR" Gate model; **A-CDM** = Additive cognitive diagnostic model; **mDINA** = multilevel DINA; **GPCM** = Generalized partial credit model.

Table 13: Overview of cognitive psychometric modelling studies of TIMSS/PISA science

Reference	Study, cycle and grade	Countries and sample size	CPM models	Q-matrix attributes / factors
Rosca (2004)	TIMSS-R G8	USA (n = 9,072)	LLTM	17 factors representing science content areas and cognitive demands from the TIMSS science framework, aspects of text structure, and properties of the key and distractor options.
Liu & McKeough (2005)	TIMSS 1995 G4/G8	USA (n = 32,922)	Partial credit Rasch model	Five attributes relating to energy concepts.
Schwab (2007)	PISA 2003 G9	Australia, Canada, Ireland, New Zealand, UK and USA (total n = 34,109)	RCML	Nine attributes reflecting the three process, content, and situation competencies of the PISA framework.
Zhang (2013)	TIMSS-R G8	USA (n = 9,072)	M-IRT	Three linguistic features; frequency of logical operator (e.g. and, or, if), number of intentional actions, events, and particles (e.g. 'in order to', 'so that'), and the average number of syllables per word, regressed onto Rasch item difficulty estimates.
Kabiri et al. (2017)	TIMSS 2011 G8	Iran (n = 6,029)	GDM	Eight attributes related to scientific competencies derived from a variety of assessment frameworks (TIMSS, PISA, NAEP etc.), such as 'using science in real-world situations' and 'predicting likely future events given existing evidence and scientific perceptions'.
Hu et al. (2021)	TIMSS 2011 G4	United States, Russia, Australia, Singapore, and England (sample sizes not reported)	DINA	Eight attributes; observing phenomena, describing phenomena, obtaining data, analysing data, using facts, constructing reflection, systematic use of theory, and scientific reasoning. Adapted from Yao, Guo & Neumann (2016).
Zhou & Traynor (2022)	TIMSS 2011	Australia (n = 6,146) Ontario (n = 4,568) Hong Kong (n = 3,957)	DINA	Six attributes relating to energy concepts (e.g. forms, sources, conductors and insulators etc.)
Harrison et al. (2023)	PISA 2015 G9 (field trial data)	13 countries (total n = 11,960 in mathematics)	GPCM	Six facets representing different contexts, systems (e.g. living, earth and space, physical systems), as well as competency, knowledge and depth of knowledge from the PISA 2015 science framework.

LLTM = Linear logistic test model; **RCML** = Random coefficient multinominal logit; **M-IRT** = Mixed item-response theory; **GDM** = General diagnostic model; **GPCM** = Generalized partial credit model

Table 14: Overview of cognitive psychometric modelling studies of PIRLS/PISA reading

Reference	Study, cycle and grade	Countries and sample size	CPMs	Q-matrix attributes / factors
Chen & Chen (2015)	PISA 2000 G9	UK (n = 1,029)	G-DINA	Five attributes based on reading processes; 'identifying explicit information', 'generalising main ideas', 'interpretation and explanation', 'making inferences' and 'evaluation and commenting'.
Chen & Chen (2016)	PISA 2000 G9	UK, USA, Australia and New Zealand (total n = 2,238)	DINA DINO G-DINA R-RUM A-CDM	Seven attributes, including the five attributes in Chen & Chen (2015) as well as two additional attributes - 'understanding charts and graphs' and 'expressing ideas in written forms'.
Yun (2017)	PIRLS 2011 G4	USA (n = 2,574)	DINA DINO G-DINA R-RUM A-CDM	Six attributes based on the PIRLS framework; the four reading comprehension processes involved in the test-item, whether the item required a constructed response, and whether the item related to an informational text.
Bulut, Bulut & Arikan (2022)	pearl's 2016 G4	13 countries (total n = 2,894)	EPCM	Three models focusing on item format, comprehension process and text complexity, followed by introduction of person-characteristics including gender and home language, and their interactions.
Harrison et al. (2023)	PISA 2015 G9 (field trial data)	13 countries (total n = 11,960 in mathematics)	GPCM	Three facets (situation, text format, and aspect) broken down into their respective levels (four, four and three levels respectively).

G-DINA = Generalized DINA; **DINA** = Deterministic Inputs Noisy "AND" Gate model; **DINO** = Deterministic Inputs Noisy "OR" Gate model; **R-RUM** = Reduced reparametrized unified model; **A-CDM** = Additive cognitive diagnostic model; **EPCM** = Explanatory partial credit model

There were also stark differences in the use of models across the three subjects. In mathematics and reading, the most commonly used CPMs were those belonging to the family of DINA models, though several of the earliest CPM studies in mathematics used the RSM. The RSM, as well as and most of the DINA models, can be described as descriptive, person-focused models. That is, the results of their analyses primarily focus on the cognitive skills of examinees, and do so by classifying the examinees into groups. By contrast, very few articles adopted item-focused CPMs, such as the linear logistic test model (LLTM) and other approaches related to the Rasch/1PL family of models. There was no specific CPM that had been commonly used to analyse data from PISA or TIMSS science, though the models that had been used were generally more item-focused than person-focused, in contrast to the mathematics-focused literature.

4.4 Discussion of articles

As previously mentioned, articles concerned with TIMSS mathematics comprise most of the CPM analyses of ILSA data. While Table 12 lists a variety of CPMs that have been used in analysing TIMSS mathematics data, the majority of these can be divided into two main categories; studies using Tatsuoka's 1983 rule-space method (RSM) rule-space method (RSM), and studies using the deterministic input, noisy ‘and’ gate (DINA) model, and/or other models within the wider DINA framework (de la Torre, 2009). While many of the earliest articles used the RSM, no articles have used the model since 2012, and within this literature it has seemingly been replaced by the DINA family of models. The following subsections provide a synoptic overview of the main CPMs used in this body of research, outlining how the models have been applied, and what findings have been made.

4.4.1 The rule-space method

Kikumi Tatsuoka, who developed the RSM, was the first to apply the method to TIMSS data. Focusing on the grade 8 TIMSS-R mathematics assessment, Tatsuoka et al. (2004) used the RSM to compare mathematics achievement across 20 of the participating countries, including a group of the six highest-performing countries, a group of English-speaking countries with similar cultural backgrounds to the United States, and groups representing European and Middle

Eastern and Islamic countries. Twenty-three attributes were included in the Q-matrix, and were developed in collaboration with a team of mathematics educators and academics involved in educational measurement and with experience teaching either school-level or college-level mathematics/statistics. These attributes were divided into three main categories; the five main content areas in the eighth-grade TIMSS mathematics assessment, nine process attributes such as ‘rule-application', ‘logical reasoning' and ‘visualising and reading figures and graphs', and finally, nine item-type attributes, such as ‘recognising patterns and sequences', ‘approximation/estimation' and ‘proportional reasoning'. All 163 items in the eighth-grade mathematics assessment were coded by three raters, though their final analysis only focused on four booklets of items due to the distributions of attributes across the booklets.

Attribute mastery probabilities were calculated for each of the attributes for each country, and used to draw international comparisons. For example, the authors noted that the attribute mastery probabilities associated with ‘evaluating and verifying options', ‘translating word expressions into equations', and ‘using information from figures and tables' were generally high across countries, whilst the attribute with the lowest overall mastery probability was in ‘using inductive thinking skills'. The top-performing countries in TIMSS-R, and in particular, Japan, had notably higher attribute mastery probabilities in many of the higher-level cognitive processing attributes than other countries. Strengths and weaknesses were also identified in specific content domains; for example, students in the United States had particularly low levels of attribute mastery in geometry relative to their overall performance, whereas this was a relative strength for students in Russia and Italy. Additionally, students in the top 11 countries all had relatively similar levels of attribute mastery on items in the whole numbers and fractions and decimals content domains, but the top four countries, all East Asian, showed higher attribute mastery probabilities in algebra, particularly over countries such as Finland, Italy, and England.

Five other articles using the RSM were located in the search, all of them using the 23 attribute framework developed by Tatsuoka. For example, Birenbaum et al. (2005) used the RSM to compared the attribute mastery profiles of eighth grade students in Singapore, the United States, and Israel. The RSM showed that, whilst students in Singapore tended to have higher proportions

of attribute mastery, the differences were most pronounced for the 9 process attributes, and smallest for the content attributes. The authors noted that students in the United States and Israel did not necessarily have much less content knowledge than students in Singapore, but had much lower proportions of mastery in higher-level "mathematical thinking" skills, particularly in a few of the attributes including "logical thinking" and "data and procedure management". In a similar study, Dogan and Tatsuoka (2008) compared the attribute mastery profiles of students in the United States to those in Turkey. It was found that students in the United States had higher levels of attribute mastery in 17 of the 23 attributes. However, Dogan and Tatsuoka conducted further analyses by creating five ‘composite attributes'; one for algebra, one for geometry, one for numbers, one for word problems, and one for ‘advanced problem solving and thinking skills', where an examinee was classified as having mastered a composite attribute if they had mastered all of the sub-attributes they were comprised of. Through analysing the most common patterns of attribute mastery, Dogan and Tatsuoka proposed three different learning progressions that students follow, and suggested that students in Turkey were more likely to master skills in geometry before those in number, whereas the inverse was true in the United States. The remaining three articles using the RSM were all conducted by Chen et al. (2006; 2010; 2011) and focused on data from the Taiwanese cohort in TIMSS-R, with all three focusing on differences in attribute mastery between male and female students, as well as those from rural and urban backgrounds. While gender differences tended to be minimal, for almost all the attributes, Taiwanese students from urban backgrounds had higher levels of mastery than students from rural backgrounds.

4.4.2 The deterministic inputs, noisy ‘and’ gate model

The DINA and the wider collection of related models make up the bulk of the articles identified in this review. The first of these was conducted by Lee et al. (2011), who used the original DINA model to analyse the mastery of different mathematical content areas in the fourth-grade TIMSS mathematics assessment. Their analysis drew comparisons between two benchmarking participants, the US states of Massachusetts and Minnesota, to the performance of the larger United States cohort. Using the TIMSS 2007 mathematics framework as a base, 15 different

attributes, each linked to one of the three content domains in TIMSS fourth-grade mathematics (number, geometric shapes and measures, and data display) were included in the Q-matrix, and were applied to 25 test-items. By comparing the DINA parameter estimates and attribute mastery among the three cohorts, the study aimed to explore whether differences in overall TIMSS mathematics scores could be attributed to any specific differences in attribute mastery.

Massachusetts and Minnesota both scored significantly higher than the United States on the TIMSS scale, and this was reflected in the proportion of correct responses given to this sample of items. However, Lee et al. (2011) highlighted the discrepancy between this, and the calculated proportion of attributes mastered by examinees in each cohort; although Massachusetts had the highest proportion of correct answers to these test-items, the average proportion of attribute mastery was lower than in the other two cohorts. In combination with the calculated guessing and slip parameters produced with the basic DINA model, it was suggested that students in Massachusetts scored well overall because they were more successfully able to guess the correct responses to test-items for which they had not mastered the required attributes, particularly in the ‘data display' content domain. Conversely, students in the United States cohort had particularly large slip parameters on four items; this included two items in the ‘number' content domain (with entirely different attributes), as well as two items in the ‘geometric shapes and measures' content domain both requiring the same attribute related to the calculation/estimation of perimeter, area and/or volume. Follow-up logistic regressions were used to compare the significance of attribute mastery differences between the three cohorts. Students in the United States had significantly higher mastery of two attributes, relating to calculation/estimation of perimeter area and/or volume, and reading data from tables/pictographs/ bar charts/pie charts, though significantly lower mastery of three attributes in the ‘number' content domain, and one in each of the other content domains.

Lee et al.’s (2011) initial use of the DINA model to assess attribute mastery in TIMSS has been used as the basis for further research focusing on cognitive diagnostic modelling of TIMSS mathematics. For example, Hansen et al. (2014) followed up on Lee et al.’s work by suggesting minor changes to the Q-matrix to improve model fit, and to better account for local dependence

between two used in the analysis. Some studies have also used the basic DINA model to assess the attributes of the TIMSS 2007 fourth-grade mathematics framework, utilising different test-items and focusing on different national samples (Chen, 2013). Other studies meanwhile have used simplified Q-matrices with a smaller number of attributes based on the content domains and/or cognitive sub-competencies in the TIMSS frameworks (Hou et al., 2013; George & Robitzsch, 2018).

However, other studies have suggested that the basic DINA model may not be the most appropriate for retrofitting to TIMSS. Yamaguchi & Okada (2018) found that although CPM models generally fit the TIMSS dataset better than IRT models such as the 1-3 parameter logistic (1-3PL) models, both the DINA model and the ‘deterministic inputs noisy ‘or; gate (DINO) model were outclassed by alternative sub-models in the wider family of DINA models, such as the A-CDM, LLM, and R-RUM models (de la Torre, 2012) that have looser rules for parameterisation.

While many of the articles using DINA models have only used descriptive variants, explanatory variants of the DINA model have also been used to explain attribute mastery as a function of other person properties. For example, Park & Lee (2014) utilised the reparameterized DINA (RDINA) model to explain attribute mastery in mathematics by way of students' performance in TIMSS science. Park & Lee first explored model fit with the RDINA model only, and then ran two further RDINA analyses that included examinee performance in TIMSS science as a covariate; one analysis specifying science ability to affect attribute mastery, and the other analysis specifying science ability to affect the mathematics items. The guessing and slip parameters for each item were relatively consistent when using the RDINA with no covariates, or when science ability was used as a covariate for attribute mastery, but guessing parameters reduced and slipping parameters increased substantially under the RDINA model with science ability as a covariate for test-items. Additionally, stronger performance in TIMSS science was positively correlated with the probability of mastery of all but one of the seven attributes in the Q-matrix (lines and angles) and had particularly large effects on three attributes (whole numbers, fractions and decimals, and location and movement).

Five articles were found to have focused on data from reading assessments, with three making use of CPMs in the DINA family. Chen and Chen conducted two studies that utilised PISA data, and adopted similar statistical approaches through their use of the generalised DINA (G-DINA) model. In both articles, the authors specified that the G-DINA model was chosen due to its relaxed assumptions about the relationship between attribute mastery and the probability of success on items that assess multiple attributes. In the first of these two articles, Chen and Chen (2015) focused on 1,029 British students who sat a collection of 20 reading items from the original PISA 2000 study. The authors, with reference to the PISA framework, a variety of reading comprehension taxonomies, and in collaboration with six literacy experts, developed an attribute framework and subsequent Q-matrix based around five reading comprehension skills. After assessing the overall model-fit, they calculated estimates of the prevalence of attribute mastery, and of the 32 (2^5) potential latent classes of attribute mastery. The five attributes ranged in mastery prevalences between 0.678 (Identification of explicitly stated information) and 0.526 (Evaluation and comments on text and author's opinions/intentions). This relatively narrow span of mastery prevalences was somewhat reflected in the most common latent classes of attribute mastery; the four most common classes accounted for more than 70% of the posterior probabilities, and this included mastery of all five attributes (the most common class, representing 25% of the student cohort), and mastery of none of the attributes (the third most common class, representing just under 16% of the students).

Chen and Chen (2016) expanded on this work by focusing on a different sample of items from PISA 2000, and by expanding the analysis to data from students in the USA, Australia, and New Zealand, to include two additional attributes, and to assess model fit under a wider variety of models in the DINA family. In terms of model fit, the basic DINA and DINO models were outperformed by the saturated models that account for the interactions between different reading comprehension skills. Through using the G-DINA model, it was also found that just over 30% of students had mastered all seven attributes, though the second most common latent class (17.8% of students) had mastery of just a single attribute, ‘evaluating and commenting', which the authors

suggested may be evidence that this attribute is more independent than the other comprehension skills, as only one other latent class had a posterior probability above 3%.

Yun's (2017) doctoral thesis focused on PIRLS, which assesses reading comprehension in the fourth grade. Six attributes were included in the formation of Q-matrix, covering the four comprehension processes included in the PIRLS 2011 framework, and two item-type attributes for constructed-response items and for informational text items (items in response to informational, as opposed to literary texts). Of the five chosen CPMs, the G-DINA model was found to have significantly better model-fit than other models. There was very little variation in attribute mastery across the four different comprehension processes (around 75% for all four), whereas mastery of constructing written responses, and in answering questions related to informational texts was lower (both around 55% attribute mastery probabilities). However, given that this analysis only focused on items clustered on one literary text, and one informational text, it is arguably not appropriate to draw conclusions about mastery of informational reading comprehension, as it was noted that the informational text scored higher in indices of text complexity.

4.4.3 Other models

Only a small number of articles utilised models other than the RSM or those in the DINA family of models. However, while these *person-focused* CPM approaches dominated research in mathematics and reading, the models used in science are notably more *item-focused*. For example, Rosca (2004) explored how the LLTM could be used to assess the difficulty of eighth-grade science test-items in the 1999 TIMSS-R study. Focusing exclusively on the 104 multiple-choice test-items used in the assessment, item difficulty was calculated based on 17 item-level features. These features were broadly concerned with different aspects of text structure and Flesch-Kincaid readability indices, TIMSS science content areas (e.g. biology, chemistry, Earth science), cognitive demands (including cognitive levels, such as knowing vs. evaluation), as well as text characteristics of the key and distractor options. Although all but one of the 17 item-level features was found to significantly contribute to item difficulty, the contribution of each feature

was small, and this was reflected in the relatively poor correlation between item difficulty estimates produced by the LLTM and by the basic Rasch model. A variety of other Rasch-based models have been applied to data from TIMSS mathematics and science (e.g. Liu & McKeough, 2005; Zhang, 2013) but these are yet to provide a cohesive overview of item difficulty in these subjects.

4.5 Concluding remarks

This review intended to provide an overview of how cognitive psychometric models have been used to analyse data from three of the largest and most influential international assessments; PISA, TIMSS, and PIRLS. Critics of these ILSAs have argued that, despite their claims, the country-level scores tell us very little about what students know or can do in the subjects they assess, and consequently, what countries can do to improve the skills of their students. Further concerns have been raised about the claims that the test-items used in these assessments measure what they purport to, and whether it can be trusted that these test-items measure student abilities invariantly across countries and language-groups. CPMs could allow for the evaluation of these concerns and allow us to better contextualise the results reported by these studies. To my knowledge however, there has been no synthesis of the body of CPM analyses of ILSA data.

Forty-two relevant articles were located. The overwhelming majority of these focused on person-diagnostic modelling of data from TIMSS mathematics. In contrast, there were very few articles focused on PISA mathematics, articles focused on science and reading assessment, or articles focused on item difficulty modelling.

Much of the earlier CPM research was based around Tatsuoka et al.'s (2004) RSM analysis of data from the eighth-grade cohort of the TIMSS-R mathematics study, using a framework of 23 attributes representing mathematical content areas, wider cognitive processing skills, and attributes reflecting the types of test-items used. The RSM analyses used identified common attribute mastery profiles of students in the strongest performing countries relative to those in countries performing near and below the middle of the TIMSS scale. More recent articles

meanwhile have typically employed a variety of models in the DINA family, which also produce attribute mastery profiles for students in the participating countries, but also incorporate guessing and slip parameters that help to evaluate the model's fit to the data, particularly in evaluating the mastery frameworks used.

Similarly to how analyses using the RSM typically expanded upon the work of one previous study, many of the DINA-based analyses located built off a single preliminary study using the model - Lee et al.'s (2011) comparisons of performance in the fourth-grade mathematics assessment of TIMSS 2007 between the whole USA cohort and the benchmarking samples from Massachusetts and Minnesota. Despite Massachusetts having a significantly higher score in TIMSS 2007 than Minnesota or the wider USA sample, Lee et al. found that this was not explained by greater mastery of any specific topics of the TIMSS mathematics framework. Rather, the DINA analysis suggested that students in Massachusetts were actually more successful at guessing the correct answers than students in Minnesota or the USA cohort. However, this conclusion relies on the assumption that the TIMSS item framework used in the attribute framework adequately accounts for what has been measured by the test-items used in TIMSS.

More broadly, concerns have been raised about the retrofitting of person-diagnostic CPMs like the DINA model to assessment datasets that were not designed with attribute mastery in mind (e.g. Ravand & Baghaei, 2020; Gierl & Cui, 2008; Williamson, 2023). While these concerns have not been explicitly tied to international assessments, it is perhaps of concern that very few articles located in this review employed CPM approaches dedicated to assessing items. Rather, there has been a somewhat overwhelming focus person-focused approaches like the RSM and particularly in recent years, CPMs within the DINA family. It is possible that the relationship between attributes and items have been assumed in these analyses because the attributes in some of the more recent studies have been directly taken from the TIMSS mathematics assessment frameworks. However, because the items in these studies were not designed with CPMs in mind, the strength of this relationship at the item-level should not be assumed.

There remains a need for further research employing CPMs with greater item-explanatory power that can evaluate the properties of the test-items rather than on the cognitive profiles of examinees. Establishing that the attributes used in these person-diagnostic analyses actually relate to the functioning of items would help justify their selection for explaining student performance. If the relationship between attributes and items is weak, there is little justification for using them to describe student performance. As it stands, this thematic review suggests that this necessary evidence is sorely overlooked in many of the existing person-diagnostic analyses.

4.5.1 Limitations and future research

It must be acknowledged that there are a few limitations of this review. Firstly, any articles that were not published in English were omitted from this review. This is despite some articles having English-language titles and/or short abstracts that indicate they might have been relevant to this review. Additionally, the search terms used were focused on English-language names of the ILSAs, though some countries have their own variant names - for example, the acronym ‘IGLU' is used in Germany, rather than PIRLS. While it is anticipated that articles exclusively using these national variants would also not be published in English, it is possible that this approach has underestimated the number of articles that could have been included.

A further limitation is that all the articles were read, categorised, and interpreted by a single researcher. The determination of which articles were of relevance to this review, and those which were not, is a somewhat subjective process, especially given the wide variety of statistical approaches that might fall under the umbrella of ‘cognitive psychometric modelling’. Related to this was the decision to exclude the large number of articles concerned with DIF-based analyses (95) from the list of relevant articles. These analyses do provide evidence that has been used to evaluate the assumptions of test invariance that these ILSAs rely on, albeit in a descriptive and somewhat limited way.

For researchers interested in expanding upon the results provided by the OECD and IEA, the current body of CPM research has shown how models, particularly those in the DINA family, can be used to provide more detailed information on the common cognitive profiles that students at

these ages possess, particularly in mathematics. Such analyses could be used to develop theoretical learning progressions in those subjects. There is also the potential for researchers to conduct more explanatory analyses that assess the contributions of other person-level factors to the mastery of different cognitive attributes. For researchers more interested in *evaluating* ILSAs, this review suggests that there is still a great amount of untapped potential for analyses targeting the properties of the test-items used in these assessments. This is an important pre-cursor to person-diagnostic analyses that has been undervalued in the existing body of literature.

PISA, TIMSS and PIRLS all boast large, nationally representative samples of students, and have the potential to provide insightful overviews of the academic abilities of students in participating countries, as well as meaningful contributions to those countries about the actions that can be taken to improve educational standards. Thorough evaluations of what drives the difficulty of the test-items, and more explanatory analyses of the cognitive profiles of the examinees can help to maximise the benefits and utility of ILSAs, but as shown in this review, these kinds of analyses remain fairly limited.

4.6 References

Addey, C. (2017). Golden relics & historical standards: how the OECD is expanding global education governance through PISA for Development. *Critical Studies in Education, 58:3*, 311-325, https://doi.org/10.1080/17508487.2017.1352006

Birenbaum, M., Tatsuoka, C., & Xin, T. (2005). Large-scale diagnostic assessment: comparison of eighth graders' mathematics performance in the United States, Singapore and Israel. *Assessment in Education: Principles, Policy & Practice, 12(2)*, 167-181. https://doi.org/10.1080/09695940500143852

Bulut, H.C., Bulut, O., & Arikan, S. (2022). Evaluating group differences in online reading comprehension: The impact of item properties. *International Journal of Testing, 23(1)*, 10-33. https://doi.org/10.1080/15305058.2022.2044821

Burkes, L.L. (2009). *Identifying differential item functioning related to student socioeconomic status and investigating sources related to classroom opportunities to learn.* Doctoral dissertation, University of Delaware. Retrieved from: https://www.proquest.com/dissertations-theses/identifying-differential-item-functioning-related/docview/304879353/se-2

Casabianca, J.M., & Lewis, C. (2018). Statistical equivalence testing approaches for Mantel–Haenszel DIF analysis. *Journal of Educational and Behavioral Statistics, 43(4)*, 407-439. https://doi.org/10.3102/1076998617742410

Chen, C.M. (2013). *Examining uncertainty and misspecification of attributes in cognitive diagnostic models.* Doctoral dissertation, Columbia University. https://doi.org/10.7916/D8PC38K3

Chen, H., & Chen, J. (2015). Exploring reading comprehension skill relationships through the G-DINA model. *Educational Psychology, 36(6)*, 1049–1064. https://doi.org/10.1080/01443410.2015.1076764

Chen, H., & Chen, J. (2016). Retrofitting non-cognitive-diagnostic reading assessment under the generalized DINA model framework*. Language Assessment Quarterly, 13(3)*, 218-230. https://doi.org/10.1080/15434303.2016.1210610

Chen, Y.H. (2011). Cognitive diagnosis of mathematics performance between rural and urban students in Taiwan. *Assessment in Education: Principles, Policy & Practice, 19(2)*, 193-209. https://doi.org/10.1080/0969594X.2011.560562

Chen, Y.H., Ferron, J.M., Thompson, M.S., Gorin, J.S., & Tatsuoka, K.K. (2010). Group comparisons of mathematics performance from a cognitive diagnostic perspective. *Educational Research and Evaluation, 16(4)*, 325-343. https://doi.org/10.1080/13803611.2010.527760

Chen, Y.H., Gorin, J., Thompson, M., & Tatsuoka, K. (2006). *Verification of cognitive attributes required to solve the TIMSS-1999 mathematics items for Taiwanese students*. Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, Apr 2006). Retrieved from: https://eric.ed.gov/?id=ED491509

Choi, K.M., Lee, Y.S., & Park, Y.S. (2015). What CDM can tell about what students have learned: an analysis of TIMSS eighth grade mathematics. *Eurasia Journal of Mathematics, Science and Technology Education, 11(6)*, 1563-1577. https://doi.org/10.12973/eurasia.2015.1421a

de la Torre, J. (2009). DINA model and parameter estimation: a didactic. *Journal of Educational and Behavioral Statistics, 34(1*), 115-130. https://doi.org/10.3102/1076998607309474

de la Torre, J. (2012). The generalized DINA model framework. *Psychometrika, 76*, 179-199. https://doi.org/10.1007/s11336-011-9207-7

Delafontaine, J., Chen, C., Park, J.Y., & Van den Noortgate, W. (2022). Using country-specific Q-matrices for cognitive diagnostic assessments with international large-scale data. *Large-scale Assessments in Education, 10(1)*, Article 19, https://doi.org/10.1186/s40536-022-00138-4

Dogan, E., Tatsuoka, K. (2008). An international comparison using a diagnostic testing model: Turkish students' profile of mathematical skills on TIMSS-R. *Educational Studies in Mathematics, 68,* 263-272. https://doi.org/10.1007/s10649-007-9099-8

El Masri, Y.H., & Andrich, D. (2020). The trade-off between model fit, invariance, and validity: the case of PISA science assessments. *Applied Measurement in Education, 33(2)*, 174-188. https://doi.org/10.1080/08957347.2020.1732384

Fischer, G.H. (1972). *Conditional maximum-likelihood estimations of item parameters for a linear logistic test model* (Research Bulletin 9). Vienna: University of Vienna, Psychological Institute. https://doi.org/10.13140/RG.2.2.36074.72647

Fischman, G.E., Topper, A. M., Silova, I., Goebel, J., & Holloway, J.L. (2018). Examining the influence of international large-scale assessments on national education policies. *Journal of Education Policy, 34(4)*, 470-499. https://doi.org/10.1080/02680939.2018.1460493

George, A.C., & Robitzsch, A. (2018). Focusing on interactions between content and cognition: a new perspective on gender differences in mathematical sub-competencies. *Applied Measurement in Education, 31(1)*, 79-97. https://doi.org/10.1080/08957347.2017.1391260

Gierl, M.J., & Cui, Y. (2008). Defining characteristics of diagnostic classification models and the problem of retrofitting in cognitive diagnostic assessment. *Measurement: Interdisciplinary Research and Perspectives, 6(4),* 263-268. https://doi.org/10.1080/15366360802497762

Gorur, R. (2017). Statistics and statecraft: exploring the potentials, politics and practices of international educational assessment. *Critical Studies in Education, 58(3)*, 261-265. https://doi.org/10.1080/17508487.2017.1353271

Hansen, M., Cai, L., Monroe, S., & Li, Z. (2014). Limited-information goodness-of-fit testing of diagnostic classification item response theory models. *British Journal of Mathematical and Statistical Psychology, 69, 3,* 225-252. https://doi.org/10.1111/bmsp.12074

Harrison, S., Kroehne, U., Goldhammer, F., Lüdtke, O., & Robitzsch, A. (2023). Comparing the score interpretation across modes in PISA: an investigation of how item facets affect difficulty. *Large-scale Assessments in Education, 11(1)*, Article 8. https://doi.org/10.1186/s40536-023-00157-9

Hou, L., Nandakumar, R., & de la Torre, J. (2013). Differential item functioning assessment in cognitive diagnostic modeling: application of the Wald test to investigate DIF in the DINA model. *Journal of Educational Measurement, 51*, 98-125. https://doi.org/10.1111/jedm.12036

Kabiri, M., Ghazi-Tabatabaei, M., Bazargan, A., Shokoohi-Yekta, M., & Kharrazi, K. (2016). Diagnosing competency mastery in science: an application of GDM to TIMSS 2011

data. *Applied Measurement in Education, 30(1)*, 27-38. https://doi.org/10.1080/08957347.2016.1258407

Lan, M.C. (2014). *Exploring gender differential item functioning (DIF) on eighth grade mathematics items for the United States and Taiwan*. Doctoral dissertation, University of Washington. Retrieved from: https://www.proquest.com/dissertations-theses/exploring-gender-differential-item-functioning/docview/1622569346/se-2

Lee, Y.S., Park, Y.S., & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the U.S. national sample using the TIMSS 2007. *International Journal of Testing, 11(2)*, 144-177. https://doi.org/10.1080/15305058.2010.534571

Liu, X. & McKeough, A. (2005). Developmental growth in students' concept of energy: analysis of selected items from the TIMSS database. *Journal of Research in Science Teaching, 42*: 493-517. https://doi.org/10.1002/tea.20060

Ma, L. (2014). *Validation of the item-attribute matrix in TIMSS-mathematics using multiple regression and the LSDM*. Doctoral thesis, University of Denver. Retrieved from: https://www.proquest.com/dissertations-theses/validation-item-attribute-matrix-timss/docview/1525999295/se-2

Ma, W. & de la Torre, J. (2016). A sequential cognitive diagnosis model for polytomous responses. *British Journal of Mathematical and Statistical Psychology, 69,* 253-275. https://doi.org/10.1111/bmsp.12070

Ma, W., & de la Torre, J. (2019). Category-level model selection for the sequential G-DINA model. *Journal of Educational and Behavioral Statistics, 44(1),* 45-77. https://doi.org/10.3102/1076998618792484

Ma, W. & de la Torre, J. (2020a). An empirical Q-matrix validation method for the sequential generalized DINA model. *British Journal of Mathematical and Statistical Psychology, 73*, 142-163. https://doi.org/10.1111/bmsp.12156

Park, Y.S., & Lee, Y.S. (2014). An extension of the DINA model using covariates: examining factors affecting response probability and latent classification. *Applied Psychological Measurement, 38(5)*, 376-390. https://doi.org/10.1177/0146621614523830

Ravand, H., & Baghaei, P. (2020). Diagnostic classification models: recent developments, practical issues, and prospects. *International Journal of Testing, 20(1),* 24-56. https://doi.org/10.1080/15305058.2019.1588278

Rosca, C.V. (2004). *What makes a science item difficult? A study of TIMSS-R items using regression and the linear logistic test model*. Doctoral dissertation, Boston College. Retrieved from: https://www.proquest.com/dissertations-theses/what-makes-science-item-difficult-study-timss-r/docview/305214920/se-2

Rupp, A.A. (2007). The answer is in the question: A guide for describing and investigating the conceptual foundations and statistical properties of cognitive psychometric models. *International Journal of Testing, 7(2)*, 95-125. https://doi.org/10.1080/15305050701193454

Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement, 74(1)*, 31-57. https://doi.org/10.1177/0013164413498257

Sandilands, D., Oliveri, M.E., Zumbo, B.D., & Ercikan, K. (2013). Investigating sources of differential item functioning in international large-scale assessments using a confirmatory approach. *International Journal of Testing, 13(2)*, 152-174. https://doi.org/10.1080/15305058.2012.690140

Schwab, C.J. (2007). *What can we learn from PISA? Investigating PISA's approach to scientific literacy*. Doctoral dissertation, University of California, Berkeley. Retrieved from: https://www.proquest.com/dissertations-theses/what-can-we-learn-pisa-investigating-pisas/docview/304899441/se-2

Skaggs, G., Wilkins, J.L.M., & Hein, S.F. (2016). Grain size and parameter recovery with TIMSS and the general diagnostic model. *International Journal of Testing, 16(4)*, 310-330. https://doi.org/10.1080/15305058.2016.1145683

Stubbe, T.C. (2011). How do different versions of a test instrument function in a single language? A DIF analysis of the PIRLS 2006 German assessments. *Educational Research and Evaluation, 17(6)*, 465-481. https://doi.org/10.1080/13803611.2011.630560

Su, Y.L. (2013). *Cognitive diagnostic analysis using hierarchically structured skills*. Doctoral thesis, University of Iowa. Retrieved from: https://www.proquest.com/dissertations-theses/cognitive-diagnostic-analysis-using/docview/1417050106/se-2

Şen, S., & Arican, M. (2016). A diagnostic comparison of Turkish and Korean students' mathematics performances on the TIMSS 2011 assessment. *Journal of Measurement and Evaluation in Education and Psychology, 6(2)*. Retrieved from: https://dergipark.org.tr/en/pub/epod/issue/27317/287594

Tatsuoka, K.K. (1983). Rule space: an approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20(4)*, 345-354. https://doi.org/10.1111/j.1745-3984.1983.tb00212.x

Tatsuoka, K K., Corter, J.E., & Tatsuoka, C. (2004). Patterns of diagnosed mathematical content and process skills in TIMSS-R across a sample of 20 countries. *American Educational Research Journal, 41(4)*, 901-926. https://doi.org/10.3102/00028312041004901

Terzi, R., & Sen, S. (2019). A nondiagnostic assessment for diagnostic purposes: Q-matrix validation and item-based model fit evaluation for the TIMSS 2011 assessment. *Sage Open, 9(1)*. https://doi.org/10.1177/2158244019832684

Toker, T. (2010). *Cognitive diagnostic assessment of TIMSS-2007 mathematics achievement items for 8th graders in Turkey*. Master's thesis, University of Denver. Retrieved from: https://www.proquest.com/dissertations-theses/cognitive-diagnostic-assessment-timss-2007/docview/759976182/se-2

Toker, T. (2016). *A comparison of latent class analysis and the mixture Rasch model: A cross-cultural comparison of 8th grade mathematics achievement in the fourth international mathematics and science study (TIMSS-2011)*. Doctoral dissertation, University of Denver. Retrieved from: https://www.proquest.com/dissertations-theses/comparison-latent-class-analysis-mixture-rasch/docview/1830757654/se-2

Toker, T., & Green, K. (2012). *An application of cognitive diagnostic assessment on TIMMS-2007 8th grade mathematics items*. Paper presented at the annual meeting of the American Educational Research Association (Vancouver, British Columbia, Canada, April 14-16, 2012). Retrieved from: https://eric.ed.gov/?id=ED543803

von Davier, M. (2020). TIMSS 2019 scaling methodology: item response theory, population models, and linking across modes. In: *Methods and Procedures: TIMSS 2019 Technical Report*, 11.1-11.25. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Wagemaker, H. (2020). Study design and evolution, and the imperatives of reliability and validity. In: *Reliability and validity of international large-scale assessment: understanding IEA's comparative studies of student achievement*. Cham: Springer International Publishing.

Walzebug, A. (2014). Is there a language-based social disadvantage in solving mathematical items*? Learning, Culture and Social Interaction, 3(2)*, 159-169. https://doi.org/10.1016/j.lcsi.2014.03.002

Wang, W.C., & Qiu, X.L. (2019). Multilevel modeling of cognitive diagnostic assessment: the multilevel DINA example. *Applied Psychological Measurement, 43(1)*, 34-50. https://doi.org/10.1177/0146621618765713

Williamson, J. (2023). *Cognitive diagnostic models and how they can be useful.* Cambridge University Press & Assessment. https://doi.org/10.17863/CAM.110853

Yamaguchi, K., & Okada, K. (2018). Comparison among cognitive diagnostic models for the TIMSS 2007 fourth grade mathematics assessment*. PloS One, 13(2),* https://doi.org/10.1371/journal.pone.0188691

Yun, J. (2017). *Investigating structures of reading comprehension attributes at different proficiency levels: applying cognitive diagnosis models and factor analyses*. Doctoral dissertation, Florida State University. Retrieved from: https://www.proquest.com/dissertations-theses/investigating-structures-reading-comprehension/docview/2017341448/se-2

Zhan, P., Jiao, H. & Liao, D. (2018). Cognitive diagnosis modelling incorporating item response times. *British Journal of Mathematical and Statistical Psychology, 71*, 262-286. https://doi.org/10.1111/bmsp.12114

Zhan, P., Liao, M., & Bian, Y. (2018). Joint testlet cognitive diagnosis modeling for paired local item dependence in response times and response accuracy. *Frontiers in Psychology, 9*. https://doi.org/10.3389/fpsyg.2018.00607

Zhang, T. (2013). *The role of reading comprehension in large-scale subject-matter assessments.* Doctoral dissertation, University of Maryland. Retrieved from: https://www.proquest.com/dissertations-theses/role-reading-comprehension-large-scale-subject/docview/1432767383/se-2

Zhao, F. (2013). *Using noncompensatory models in cognitive diagnostic mathematics assessments: An evaluation based on empirical data*. Doctoral dissertation, University of

Kansas. Retrieved from: https://www.proquest.com/dissertations-theses/using-noncompensatory-models-cognitive-diagnostic/docview/1428738807/se-2

Zhou, S., & Traynor, A. (2022). Measuring students' learning progressions in energy using cognitive diagnostic models. *Frontiers in Psychology, 13*, Article 892884. https://doi.org/10.3389/fpsyg.2022.892884

5 Paper 2: Sources of item difficulty in mathematics assessment: an evaluation of the TIMSS 2011 mathematics framework

Test-items used in the Trends in International Mathematics and Science Study (TIMSS) are constructed around an assessment framework that defines the range of topics and cognitive skills that TIMSS intends to assess. This framework includes the subscales that provide more nuanced discussion of differences in performance profiles across countries, and often form the foundation of other analyses that perform cognitive diagnoses of students' mathematical competencies. What has not been well established is how well the framework actually explains differences in the psychometric difficulty of these test-items.

This paper uses item-level performance data from approximately 20,000 students in six Anglophone countries who participated in the 2011 cycle of TIMSS. The linear logistic test model (LLTM) was used to analyse how well attributes within the assessment framework predict the difficulty of its fourth-grade mathematics test-items. Major content and cognitive domains of the framework, as well as finer topic areas and item-level reading demands are incorporated into the modelling process.

Results show that attributes of the assessment framework fail to explain a large proportion of the variance in item difficulty estimates, including in comparison to an alternative proposed framework. This paper therefore raises questions about the interpretations drawn regarding performance on TIMSS mathematics subscales and other analyses that make specific inferences about students' mathematical competencies around this framework.

5.1 Introduction

In designing educational assessments, one of the key considerations is in ensuring that they validly assess the constructs they intend to measure. This is true not only at the overall test level, but also at the individual item level. Good test-items are those that are well targeted to the ability levels of the assessment population (Aldrich et al., 2024), that are fair and balanced to sub-

groups within the assessment population (e.g. are equally challenging for examinees with different characteristics unrelated to the assessment construct, such as gender, socioeconomic background or for students of different language groups; e.g. Casabianca & Lewis, 2018; Walzebug, 2014; El Masri & Andrich, 2020).

Test-items should also be predictably difficult. That is, there should be a prior understanding of what an examinee needs to know and be able to do to answer the question, and what features or attributes of the test-item contribute to that item's difficulty (Pollitt et al., 2007). This includes ensuring the test-item eliminates or minimises the impact of construct-irrelevant features that affect the difficulty of the item (Messick, 1993). The collective body of test-items used in an educational assessment should cover the range of abilities in the testing population to provide fine-grained discriminations between examinees of differing ability levels. To design test-items that meet this requirement, the empirical difficulty of a test-item *should* be reasonably predictable based on the subject content it assesses, the cognitive processes required by the examinee to produce a correct answer, and other recognised factors that affect the way examinees interact with test-items.

In practice however, it is not always easy to predict how difficult an individual test-item may be for the examinees prior to formal testing (Hambleton & Jirka, 2006; Pettersen & Nortvedt, 2018). Expectations of how difficult a test-item may be for the examinee population do not always align well with empirical measures of item difficulty (e.g. Hambleton & Jirka, 2006; Melican et al., 1989), though collective judgements made by groups of relative difficulty compared to other items have been found to be more successful in predicting actual item difficulties (e.g. Holmes et al., 2018). It has been suggested that most attempts at modelling the difficulty of test-items explain less than 20% of the variance in item difficulty estimates (El Masri et al., 2017), highlighting the challenges faced by those attempting to design test-items with predictable item difficulties.

Major international large-scale assessments (ILSAs), such as the Trends in International Mathematics and Science Study (TIMSS), Programme for International Student Assessment (PISA) and Progress in International Reading Literacy Study (PIRLS) all develop their test-items

under assessment frameworks. These outline the content coverage of their assessments, and the cognitive skills they intend to assess. For example, the International Association for the Evaluation of Educational Achievement (IEA), who conduct the TIMSS and PIRLS studies, publish assessment frameworks that describe the different content areas that their items cover. In the case of TIMSS mathematics, for example, the assessment framework identifies the major mathematics content domains assessed at the fourth and eighth grade (e.g. geometry, data), topics within those contain domains (e.g. two- and three-dimensional shapes, reading and interpreting data etc.) as well as overarching cognitive skills that examinees will need to demonstrate to answer their questions. These content and cognitive areas are directly linked to each individual item included in TIMSS, such that each item targets one content, and one cognitive domain.

Given the scale and reach of these ILSAs, there is a clear need to evaluate the validity of the assessment frameworks and how they have been applied to items; these frameworks not only inform item development, but also shape the reporting of results. Misfit between the intentions of the assessment, and the reality of what is assessed have wide implications on how the results of these studies should be interpreted.

Validating the frameworks used for designing and analysing item-level performance in these educational assessments is also important because these frameworks often serve as the basis for other research using ILSA data to more deeply evaluate student performance. A growing body of literature concerned with cognitive diagnostic analysis of student performance has made use of the datasets and assessment frameworks from ILSAs such as TIMSS to contextualise differences in performance between countries and/or sub-groups of students. For example, Lee et al. (2011) used the mathematics assessment framework from TIMSS 2007 to compare the performance of students in Massachusetts and Minnesota with those of the main USA cohort. They used 15 item-level attributes based on the fourth-grade mathematics assessment framework used in TIMSS 2007 (Mullis et al., 2005), split across the three major content domains in TIMSS – number, geometry, and data. Their analysis adopted a cognitive diagnostic modelling approach, looking at differences in the patterns of mastery and non-mastery of each of the 15 attributes. Despite

Massachusetts and Minnesota having significantly higher scores in TIMSS 2007 than the wider USA sample, Lee et al. found that this was not explained by greater mastery of any specific topics of the TIMSS mathematics framework. Rather, their analysis suggested that students in Massachusetts and Minnesota were actually more successful at guessing the correct answers than students in the wider USA cohort, particularly on items requiring examinees to read and interpret data from tables and graphs (one of the 15 attributes).

However, this conclusion relies on the assumption that the TIMSS item framework used in their analysis actually relates to how difficult the items were. While a number of other studies have built on the work of Lee et al. (2011), also making use of the TIMSS framework (e.g. Park & Lee, 2014; Ma & de la Torre, 2016; Yamaguchi & Okada, 2018; Delafontaine et al., 2022), there has been considerably less published work on whether these assessment frameworks actually relate to the difficulty of the test-items (see paper 1, Chapter 4).

When making specific comments that attempt to explain students' differential performance on items with different attributes (e.g. between geometry and data questions), there must be a strong relationship between item difficulty and those attributes. If the differences in item difficulties are poorly explained by the framework of attributes assumed to account for those differences, then conclusions drawn based on that framework are likely to be wrong. If the attributes of the TIMSS assessment framework fail to explain most of the variance in item difficulties, it threatens the validity of cognitive diagnostic modelling studies using the framework as their model for explaining student performance.

The purpose of this paper is two-fold. Firstly, to evaluate how well the TIMSS assessment framework from the 2011 cycle of the study predicts the difficulty of items in TIMSS. Second, to provide comparisons to how an alternative framework adapted from a cognitive diagnostic modelling study previously applied to TIMSS data (Tatsuoka et al., 2004) may predict the difficulty of these items. This paper conducts analyses using the one-parameter logistic model (also known as the Rasch model; Rasch, 1960) and the linear logistic test model (LLTM; Fischer,

1972) to assess the extent to which item difficulty estimates calculated using the LLTM under these frameworks correlate with item difficulty estimates that are calculated freely under the 1PL.

5.2 Methods

5.2.1 Dataset and sample

This analysis uses data from 20,319 students who participated in the TIMSS 2011 fourth-grade mathematics assessment. These data are publicly available from the TIMSS & PIRLS website (https://timssandpirls.bc.edu/). Data were taken from students in six participating countries; Australia, England, Ireland, New Zealand, Northern Ireland, and the United States. These countries were chosen because they are all Anglophone countries that had similar levels of overall mathematics performance in TIMSS 2011. While there were other countries that participated in TIMSS 2011 that also administered the assessment in English, such as Singapore and Botswana, these were countries where English was commonly not the first language of the students, that were less culturally similar to the chosen countries, and had substantially different levels of overall mathematics performance. The purpose of this study is not to draw international comparisons, but rather to use data from six countries in order to build strong sample sizes for analysis at the item level.

To ensure reading-related demands were consistent for all students in the sample, the small number of students in Ireland and Northern Ireland completed their assessments in Irish were excluded from the analysis. The final sample size of this study's sub-cohort, and average TIMSS mathematics scores of each country are presented in Table 15.

Table 15: TIMSS 2011 sample size and average performance of sub-cohort

Country	Number of examinees	Average TIMSS mathematics Score	Average 'Number' score	Average 'Geometry' score	Average 'Data' score
Australia	3,497	517 (3.2)	511 (3.5)	536 (3.2)	517 (3.6)
England	1,925	545 (3.6)	542 (3.7)	548 (4.0)	553 (4.6)
Ireland	2,511	529 (3.1)	534 (3.2)	521 (3.9)	523 (3.8)
New Zealand	3,160	486 (2.8)	484 (3.0)	484 (2.9)	491 (3.1)
Northern Ireland	2,040	563 (3.4)	566 (3.4)	560 (4.0)	556 (3.7)
United States	7,186	539 (2.2)	540 (2.3)	534 (2.5)	544 (2.0)
Total sample	**20,319**	**-**	**-**	**-**	**-**

Standard errors of performance estimates in parentheses

5.2.1.1 TIMSS 2011 booklet structure

In TIMSS 2011, not all students were administered the same mathematics test-items. Instead, each item was assigned to one of 14 blocks (numbered M01 to M14), with each block containing around 10-14 items. Students who participated in TIMSS 2011 were given one of 14 booklets, which contained two of these consecutively numbered blocks (e.g. M01 and M02, M07 and M08, M14 and M01 etc.). The reason for this approach is to ensure that the assessment as a whole targets the entire breadth of the TIMSS mathematics framework in sufficient depth, while minimising the demands placed on individual students. This approach does however introduce additional complexities into data analysis and the estimation of students' performance.

In the analysis presented in this paper, it is not possible to include data from all students and item blocks. This is because some of the item blocks used in previous cycles of TIMSS have not publicly released by the IEA because they are intended to be reused in future cycles of the study. These 'trend items' allow for equating of item parameters (including item difficulties) across different cycles of the study, allowing for time invariant reporting of country performance. To date, three of the item blocks used in TIMSS 2011 (M08, M12, and M14) have yet to be publicly released. While basic information on these items has been released (e.g. content and cognitive domains, estimated difficulty parameters), it is not possible to conduct a detailed analysis of item content for the analysis this paper requires. Given the booklet structure used in TIMSS 2011 and

the challenges that the missing blocks present on calibrating item difficulties and person ability estimates, this paper only focuses on items within the largest contiguous chain of released blocks, namely, M01 to M07. This represents approximately half of the fourth-grade mathematics test-items used in TIMSS 2011.

5.2.1.2 Test-items

A total of 85 items distributed across item blocks M01 to M07 were included in the analysis. Each student included in the analysis completed at least two items from one of these blocks, though most pupils had complete or near complete data for all items within two item blocks, and therefore each student has responses to around 25 items. Students who were given booklets 1, 2, 3, 4, 5 or 6 have usable data from two item blocks, while students given booklet 7 or 14 have usable data from a single block (M07 and M01 respectively). All students given booklets 8 – 13, or who did not have at least two valid responses to items in blocks M01 to M07 were excluded from the analysis.

TIMSS 2011 consisted of a mix of multiple-choice and constructed response items. Of the 85 items included in this analysis, 42 items were multiple-choice, while the other 43 required a constructed response. Each multiple-choice item consisted of one correct response and three distractors. Seven of the constructed response items were scored out of two, with a partially correct response being awarded one mark. For the purpose of this analysis, these items were dichotomised, such that only a fully correct response was coded as correct, while partially correct responses were recoded as incorrect. Table 16 provides a summary of the frequency of major item characteristics, including their response type, scoring type (dichotomous or polytomous), and which content and cognitive domains they focused on.

Table 16: Summary of items by their major characteristics

Major item characteristic	No. of items
Response type	
Multiple-choice	42
Constructed response	43
Scoring type	
Dichotomous	78
Polytomous	7
Content domain	
Number	46
Geometric shapes and measures	28
Data display	11
Cognitive domain	
Knowing	31
Applying	36
Reasoning	18

5.2.2 Data analysis

This section briefly covers the two main models used in this analysis; the one-parameter logistic model (1PL), and the linear logistic test model (LLTM). All main analyses were conducted in R, using the eRm package (extended Rasch modelling - Mair et al.; 2024).

5.2.2.1 One-parameter Logistic Model (Rasch Model)

The one-parameter logistic model (1PL) describes test-items in terms of a single parameter, item difficulty. The 1PL produces estimates of the difficulty of test-items, and the latent ability of each examinee. These estimates are independent of each other in that, unlike in the item or person scoring estimates produced under classical test theory models, the estimates of item difficulty produced by IRT models do not depend on the ability of the examinees. Item difficulty estimates and examinee ability estimates are however calculated on the same scale, and allow for estimations of the likelihood of an examinee's success on a test-item, as a function of the estimate of that examinee's ability relative to the difficulty of the item on that scale.

The formula for the 1PL model is as follows:

$$P\left(X_{ie}=1 \vee \theta_e, \beta_i\right)=\frac{\exp^a\left(\theta_e-\beta_i\right)}{1+\exp^a\left(\theta_e-\beta_i\right)}$$

The probability of examinee e providing a correct response to test-item i ($X_{ie}=1$¿ is a function of the examinee's ability (θ_e) and the difficulty of the item (β_i). An item's difficulty is equal to the estimated ability level of an examinee who has a 50% chance of providing the correct response. Thus, if an examinee's estimated ability level is greater than the estimated difficulty of a given item, the model suggests that there is a greater than 50% chance that the examinee provides the correct response. The 1PL assumes that all items are equally discriminating (constant value of α for all items). Other logistic models, such as the 2PL and 3PL, introduce other item-level parameters into the equation. These are an item-level discrimination parameter (a_i) and an item-level guessing parameter (c_i). Item parameter estimation in TIMSS 2011 was conducted through the use of both 2PL and 3PL models, for constructed-response and multiple-choice items respectively.

A requirement of the 1PL (as well as variants with more parameters) is that the estimates of the difficulties of a set of items in a test should be the same (within a small standard error of measurement) if administered to a set of higher performing examinees as they would be if administered to a set of lower performing examinees. Similarly, the estimates of each examinee's ability should not be dependent on the difficulty of the test-items they were given (again within a small standard error of measurement).

5.2.2.2 Linear logistic test model (LLTM)

The linear logistic test model (LLTM), initially proposed by Fischer (1972), provides an extension to the calculation of item difficulty parameters in the 1PL. Like the 1PL, the LLTM assumes unidimensionality of the test. However, the LLTM also assumes that the difficulty of each test-item is the composite of various different factors associated with that construct. In the case of a mathematics assessment, for example, the LLTM assumes that the test as a whole assesses a singular construct – mathematical ability – but that each item of the test is made more or less difficult by the different facets of mathematics that the item tests, or by other attributes of those test-items.

The LLTM imposes a restriction on the calculation of item difficulty parameters (β_i) in the 1PL formula, such that:

$$\beta_i = \sum_{j=1}^{J} q_{ij} n_j + c$$

where q_{ij} is the weight of each proposed difficulty factor j for item i, and n_j *is* the estimated difficulty parameter for each proposed difficulty factor / attribute j. A normalisation constant c is added, such that the mean of item difficulties under the LLTM is equal to 0. The 1PL can be seen as a special case of the LLTM, where the number of attributes is equal to the number of test-items (Fischer, 1972). One of the key components of the LLTM is the Q-matrix, which is established a priori to specify which item attributes are associated with each individual test-item (the q_{ij} in the LLTM formula).

The purpose of this analysis is to assess the fit between the item difficulty estimates produced under the 1PL, where a separate item difficulty parameter is calculated for each item, and the item difficulty parameters calculated under the LLTM, where each item's difficulty is the linear sum of the difficulty parameters of the attributes associated with that item. Likelihood ratio tests can be used to compare the fit of the 1PL and LLTM, and can be supported by assessing the item-level correlations between 1PL and LLTM difficulty estimates. If there is a high correlation between the item difficulty parameters produced under the 1PL and the LLTM, it suggests that the proposed Q-matrix of item attributes adequately explains the driving factors of the 1PL's estimates of item difficulty. Poor correlation, however, suggests either that the Q-matrix does not adequately account for the factors contributing to the difficulty of the test-items, or that the assignment of attributes to items in the Q-matrix has not been conducted accurately. An alternative reason may be that the LLTM's assumption that difficulty factors combine linearly may fit the data poorly.

This paper will explore model fit using four different Q-matrices of proposed difficulty factors.

5.2.3 Q-matrices

5.2.3.1 Q-matrices A-C (TIMSS framework)

In the first stage, model fit will be explored through the use of three Q-matrices; herein, Q-matrices A - C. Each of these Q-matrices have been adopted from the item categorisations and codes utilised by the IEA in their classification of the test-items used in TIMSS 2011, and therefore no new item coding was required. These three Q-matrices are outlined in Table 17.

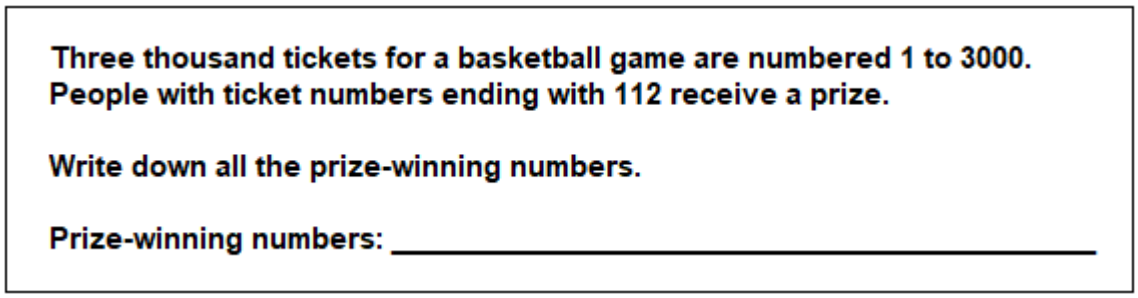
Table 17: Outline of attributes in Q-matrices A-C

Q-matrix	Summary of attributes
A	**Six attributes based on the TIMSS 2011 cognitive process and content domains.** --- **Cognitive attributes:** j_{A01} – Knowing, j_{A02} – Applying, j_{A03} – Reasoning --- **Content attributes:** j_{A04} – Number, j_{A05} – Geometry, j_{A06} – Reasoning
B	**Thirteen attributes based on the TIMSS 2011 cognitive domains, nine attributes derived from the content domains, topic areas and topic bullets, and one for whether the item requires a constructed response** --- **Cognitive attributes:** j_{B01} – Knowing, j_{B02} – Applying, j_{B03} – Reasoning --- **Content attributes:** j_{B04} – Number sense / place value awareness j_{B05} – Computations with whole numbers j_{B06} – Problem-solving with whole numbers j_{B07} – Understanding of equivalent fractions j_{B08} – Computations with fractions and decimals j_{B09} – Patterns and relationships j_{B10} – Points, lines, and angles j_{B11} – 2D and 3D Shapes j_{B12} – Data --- **Response attributes:** j_{B13} – Item requires a constructed response
C	**Fifteen attributes, the same as in Q-matrix B but also incorporating two attributes from the TIMSSPIRLS Relationships Report on reading demands.** --- **Cognitive attributes:** j_{C01} – Knowing, j_{C02} – Applying, j_{C03} – Reasoning --- **Content attributes:** j_{C04} – Number sense / place value awareness j_{C05} – Computations with whole numbers j_{C06} – Problem-solving with whole numbers j_{C07} – Understanding of equivalent fractions j_{C08} – Computations with fractions and decimals j_{C09} – Patterns and relationships j_{C10} – Points, lines, and angles j_{C11} – 2D and 3D Shapes j_{C12} – Data --- **Response attributes:** j_{C13} – Item requires a constructed response --- **Reading attributes:** j_{C14} – Item classified as having high reading demands j_{C15} – Item includes at least one use of mathematical / technical language

Q-matrix A consists of six attributes. These correspond to the three cognitive domains of the TIMSS framework and are labelled j_{A01} to j_{A03}. The other three attributes represent the three content domains of TIMSS 2011; number problems, geometry problems, and data-related problems and are labelled j_{A04} to j_{A06}. Each item belongs to one of the three cognitive domains, and one of the three content domains, and therefore, each item in the Q-matrix was defined as assessing two of the six attributes.

Q-matrix B builds on Q-matrix A by providing more detailed information about the content domain. Each item used in TIMSS 2011 was also associated with one topic area associated with its cognitive domain (e.g. 'fractions and decimals' in the number domain). Within each topic area, each item was also classified as belonging to one topic bullet, providing a more specific description of the mathematical content of the item. For example, Figure 10 shows an example item used in TIMSS 2011. It was categorised as testing the reasoning cognitive domain, the number content domain, and within this, belonged to the 'whole numbers' topic area, and 'problem solving with whole numbers' topic bullet. It was also coded as requiring a constructed response.

Figure 10: Example item 49 (M031016)



Source: TIMSS 2011 Assessment. Copyright © 2013 International Association for the Evaluation of Educational Achievement (IEA). Publisher: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Decisions about which level of specificity to include in Q-matrix B were made based on the frequencies of items within each topic bullet, then topic area, followed by content domain. Items covering the number domain were best represented among the 85 items included in the analysis, and in Q-matrix B are split across five topic bullets (j_{B04} to j_{B08}) covering whole numbers and

fractions and decimals, as well as items covering the 'patterns and relationships' topic area (j_{B09}) . Geometry items are covered by the two geometry topic areas (j_{B10} and j_{B11}) but are not further divided into topic bullets as there was not sufficient coverage of these. Further, as there were only 11 items covering the data content domain, there was no additional breakdown of data items into their respective topic areas or topic bullets (j_{B12} only).

Q-matrix C introduces two additional attributes derived from the TIMSSPIRLS 2011 Relationships Report (Martin & Mullis, 2013), which explored relationships between student performance in TIMSS and PIRLS 2011 in countries which administered both studies to the same cohort of students. In the 2011 cycle of TIMSS, the IEA made use of the overlapping cycles of their sister-study, PIRLS, to explore relationships between performances in mathematics and science with those in reading comprehension. They also recognised that reading demands in mathematics and science test-items could act as an unwanted source of difficulty in their test-items. One of their analyses focuses on reading demands in TIMSS mathematics and science items. A team of 10 coders made a holistic judgement about the reading demands in each item based off of four main aspects of the readability of the item; the total word count, the frequency of mathematical (or scientific) vocabulary in the item (e.g. multiply, kilometres, estimate), a count of symbolic language (e.g. Arabic numerals (e.g. 12, 2000), units (e.g. cm, kg), or mathematical operators (e.g. +, etc.), and the density of visual displays. From these four aspects of the item, the team rated each item as either having high, medium, or low reading demands. Q-matrix C incorporates this holistic judgement into the framework (j_{C14}) and also makes use of one of these individual criteria used within the holistic judgement – the presence of mathematically relevant technical language (j_{C15}). Q-matrix C denotes whether an item was judged to have high reading demands or not, and whether the item included at least one case of technical language.

Of the 85 items included in this analysis, 30 items were judged to have high reading demands, 25 were judged to have medium demands, and 30 were judged to have low reading demands. For the purpose of Q-matrix C, a binary distinction was made between items with high reading demands and non-high demands (i.e. medium or low). Additionally, 47 items included at least one

example of technical language. While instances of technical language formed a part of the judgement as to whether an item had high reading demands, it should be noted that only 13 of the 30 items with high reading demands included an example of technical language. However, there was a strong association between reading demand and whether the item covered the data content domain; of the 11 data items, nine were judged to have high reading demands, with one medium and one low reading demand item. The TIMSS 2011 relationships report (Martin & Mullis, 2013) notes that this is primarily due to the majority of these items including dense visual displays with many elements (e.g. different graphs and tables).

5.2.3.2 Q-matrix D

Q-matrix D was adapted from the Q-matrix first used by Tatsuoka et al. (2004) in their exploration of attribute mastery of examinees in the eighth-grade assessment of TIMSS-R 1999. A team of researchers involved in educational assessment as well as two secondary school mathematics teachers worked to develop a cognitive model for the mathematical knowledge and cognitive processes required by examinees to answer TIMSS mathematics questions. They also developed a list of more specific skills students would require to deal with items with different properties (e.g. response types, language). In total, they proposed 23 item-level attributes clustered into three groups - content attributes, process attributes, and skill attributes. This Q-matrix was subsequently used and adapted by several other articles concerned with diagnostic modelling of students' mastery of different mathematical skills (e.g. Birenbaum et al., 2005; Toker & Green, 2012; Chen, 2011), and has been selected here as a form of comparison to the IEA's assessment framework.

Given the differing ages of the cohorts, as well as changes to the content coverage of TIMSS mathematics items over time, some of the aspects in their Q-matrix were not well tailored to the items in the fourth-grade assessment in 2011. Additionally, even through the use of 85 items, around half of the total items included in the fourth-grade TIMSS 2011 mathematics assessment, some attributes were assigned to no, or very few (less than three) items. These attributes were subsequently removed or reintegrated with a similar, more well represented attribute.

The final list of attributes for Q-matrix D is reported in Table 18. Each item included in this analysis was coded to belong to one of the five content attributes, and most belonged to a single process attribute, though some items were classified as belonging to two or more. Items could be classified as testing any number of the skill attributes.

Table 18: Outline of attributes in Q-matrix D

Attribute	Description
Content	
j_{D01}	Mathematical concepts and operations involving whole numbers
j_{D02}	Mathematical concepts and operations involving fractions and decimals
j_{D03}	Geometry content
j_{D04}	Data-related content
j_{D05}	Measuring or estimating in real-world contexts: (e.g. lengths, time, temperature etc.)
Process	
j_{D06}	Translate/formulate equations and expressions
j_{D07}	Perform computations in arithmetic / geometry
j_{D08}	Make judgements or evaluate mathematical correctness
j_{D09}	Logical reasoning in written contexts
j_{D10}	Logical reasoning in visual contexts
Skill	
j_{D11}	Question requires a demonstration of number sense / place-value awareness
j_{D12}	Question requires the student to identify and understand numerical patterns and sequences.
j_{D13}	Question requires the student to identify and understand proportional relationships.
j_{D14}	Question requires the student to compare the properties of different entities in the item.
j_{D15}	Question requires interpretation of information presented in figures, tables, or charts/graphs
j_{D16}	Question requires the student to check/evaluate from a set of options (not just multiple-choice, requires active selection of most appropriate option etc.)
j_{D17}	Question requires an open-ended response
j_{D18}	Question requires verbal interpretation
j_{D19}	Question is embedded in a novel or non-routine context (usually in a real-world context)

Figure 10 (presented previously) shows an example item from TIMSS 2011. Under Q-matrix D, this item was coded as targeting content attribute j_{D01}, as it deals with whole numbers, process

attribute j_{D09}, as it is a reasoning problem embedded in a written context, j_{D11} as it tests the student's understanding of place-value, as well as skill attributes j_{D17}, j_{D18} and j_{D19} as the question has an open-ended response where a list of the three correct answers is expected, the student needs to verbally interpret what a 'prize-winning number' is from a lengthier introductory sentence, and the question is set in a novel, real-world context.

5.3 Results

5.3.1 1PL difficulty estimates

In the first step, item difficulty estimates were calculated under the 1PL model. These estimates were compared with the difficulty parameters released by the IEA, to assess the overall correlation of item difficulty estimates calculated for this sub-cohort of students, and only when including data from blocks M01 to M07. It should also be noted that the IEA used a 3PL model to calculate separate discrimination, difficulty, and pseudo-guessing parameters for multiple-choice items, and a 2PL model for items requiring a constructed response (Martin & Mullis, 2012). A partial credit model is also employed for the small number of items with polytomous scoring. The decision to use a 1PL model for all items in the analysis presented here was made to ensure compatibility with the LLTM in later analyses – a model that is a more general case of the 1PL.

The correlation between the item difficulty estimates using the sub-cohort's data and the IEA's difficulty parameters was relatively strong ($r = .81, r^2 = .65$). To assess the impact of using the 1PL for all items rather than the mix of models used by the IEA, item difficulties for each of the 85 items in this analysis were also estimated under the corresponding 2PL or 3PL model. Compared to the 1PL approach, difficulty estimates under a 2PL model for constructed response items and a 3PL model for multiple choice items produced a slightly stronger, though roughly comparable correlation with the IEA's difficulty parameters ($r = .87, r^2 = .75$). Given the relatively strong correlation with the IEA's estimates, the 1PL estimates were retained to allow for comparisons with the LLTM.

To further assess the suitability of the 1PL, outfit residual statistics were calculated for each item. Given missing data arising from the booklet structure used in TIMSS, outfit statistics were calculated separately for each item in each of the included booklets, resulting in two outfit residuals for each item. Items with at least one outfit residual greater than 1.3 were flagged as potentially problematic and assessed further through the exploration of item characteristic curves. In total, nine items had a least one outfit residual greater than 1.3. Of these, five had two outfit residuals greater than 1.3. In assessing the item characteristic curves, one item, item 13 (M051007), showed a consistent pattern across the two booklets deviating from the 1PL model. This was a multiple-choice 'number' item with an estimated difficulty of 1.21. A greater proportion of low ability examinees answered the question correctly than would be expected under the 1PL, and conversely, a considerably lower proportion of examinees with ability estimates greater than 1.21 were able to select the correct answer. This is indicative that there may have been a high level of guessing for this question, and it is likely that a 3-parameter model that estimates a pseudo-guessing parameter would better explain the functioning of this item. The impact of this item on overall model fit is likely to be very low, and so was retained in this analysis. It should however be noted that there may be an argument for the exclusion of this, and/or other items that do not adhere well to the 1PL.

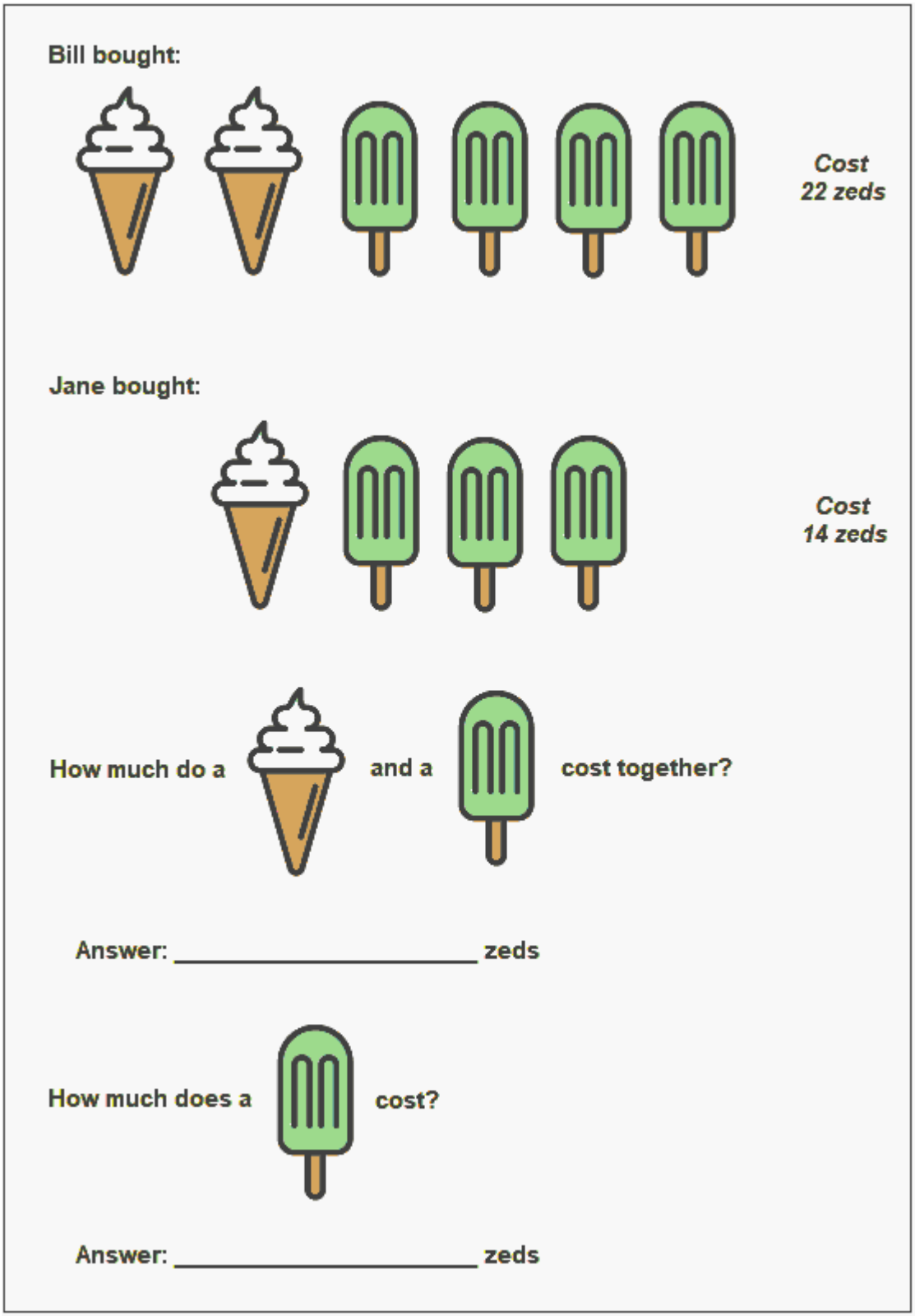
On inspecting the items with the largest rank differences in item difficulty between the sub-cohort and the IEA's difficulty parameters, there were some common trends. When ranking the 85 items in terms of their estimated difficulty, the six items which were estimated to be considerably easier for the sub-cohort were all items focused on either identifying equivalent fractions, or about identifying lines of symmetry in two-dimensional shapes. By contrast, two of the items which, by ranking, were estimated to be substantially more difficult for the sub-cohort than the wider TIMSS 2011 study cohort were items involving routine calculations involving whole numbers (e.g. 4,809 – 532). These may reflect differences in the educational experiences and curricula of these countries.

The other items with the greatest differences in ranking tended to be polytomous items that had undergone dichotomisation; as would be expected, negating the responses of partially correct

answers led to an increase in estimated difficulty. For example, item M051006 (shown in Figure 11) was estimated to be the most difficult item for the sub-cohort, with an estimated difficulty of 2.75 under the 1PL. This compared to the IEA's estimate of 1.06.

The range in item difficulty estimates was also considerably wider in this cohort than the IEA's parameters; item difficulty parameters for these items ranged between -2.74 and 2.75 for the sub-cohort, compared to just -2.26 and 1.41 in the wider cohort. This range narrowed slightly when utilising 2PL and 3PL models, though was still greater than the range of IEA estimates.

Figure 11: Example difficult item 39 (M051006)

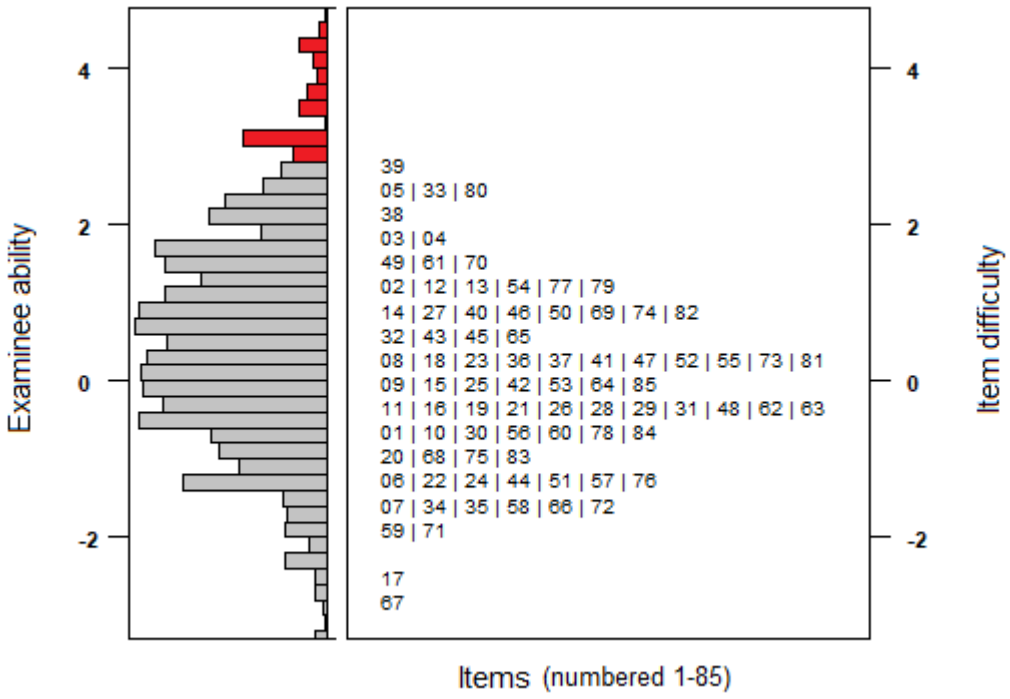


Estimated 1PL difficulty = 2.75

Source: TIMSS 2011 Assessment. Copyright © 2013 International Association for the Evaluation of Educational Achievement (IEA). Publisher: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

An item-person map was constructed to display the relationship between examinee ability estimates and the 1PL difficulty estimates of the 85 test-items, as shown in Figure 12. The group of examinees with ability estimates greater than any of the item-level item difficulty estimates are highlighted in red.

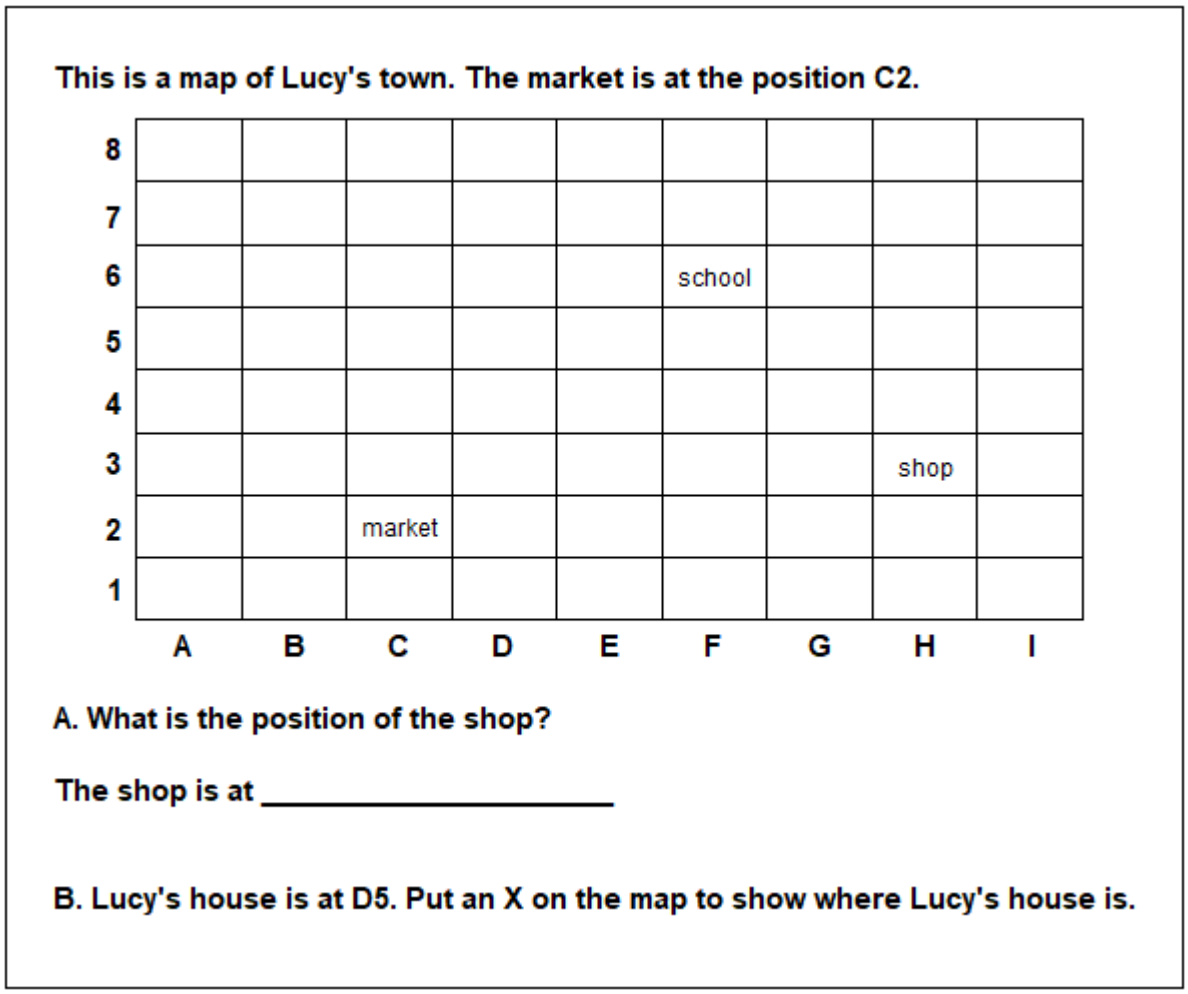
Figure 12: Item-person map of examinee ability and item difficulty distributions



The average ability level of the sub-cohort of examinees ($mean\,\theta = 0.58, SE = 0.01$) was greater than the average difficulty of the items ($mean\,\beta = 0.00, SE = 0.04$). The mean examinee ability level was significantly greater than the mean item difficulty; $t(20402) = 3.57, p < 0.001$. This means that, in general, the items in TIMSS 2011 were somewhat easy for the students in the six sub-cohort countries. However, the distribution of examinee ability levels was wide, with most items being well calibrated to the varying levels of ability; it should be reiterated that individual examinees were only given around 25 of these items (with some even less), and therefore, it was quite likely for the highest performing examinees to receive few/no items tailored to their ability. One item (number 67; M041160B, shown in Figure 13) had an estimated difficulty of -2.74, which was lower than the estimated ability of almost every examinee in the sub-cohort, suggesting that it provided little discriminatory information on the ability levels of this group of examinees. By

contrast, there were more than 1,400 examinees (around 7% of the total sub-cohort) with estimated ability levels above 2.75, highlighted in red. This is higher than that of the estimated difficulty of the hardest item, item 39 (shown in Figure 11), suggesting that the assessment is poor at discriminating between the ability levels of these examinees. It should also be reiterated that only around a quarter of examinees included in this analysis were administered this particular item, and therefore an even larger proportion of examinees may have had no items that were targeted to their ability level.

Figure 13: Example easy items 66 (M041160A) and 67 (M041160B)



Estimated 1PL item difficulty = -1.61 and -2.74 respectively.

5.3.2 LLTM difficulty estimates

5.3.2.1 Q-matrix A

Item difficulties were then calculated under the LLTM, first using Q-matrix A. Given the structure of the Q-matrix, where every item belonged to one of three cognitive attributes, and one of three content attributes, it was necessary to define reference categories to avoid multicollinearity. In this case, attributes j_{A01} (knowing) and j_{A04} (number) were selected as the reference categories.

The six attributes of Q-matrix A, and their associated eta parameters as estimated by the LLTM are reported in Table 19. All four of the entered item attributes had statistically significant eta parameters at the 95% confidence level, indicating that each of the content and cognitive domains had significantly different effects on item difficulty estimates. However, these tests are powerful given the sample sizes, and for some attributes, the size of the effect was small. For instance, there was little difference in the difficulties of 'knowing' and 'applying' items (given the very small eta parameter for 'applying', but this still represented a statistically significant difference. As would be anticipated, reasoning items were considerably more difficult, being associated with around a 1.1 logit difference in item difficulty compared to items associated with the other cognitive domains. Of the content domains, number items were slightly harder than geometry items, and considerably more difficult than data items, with data items being around 0.9 logits easier.

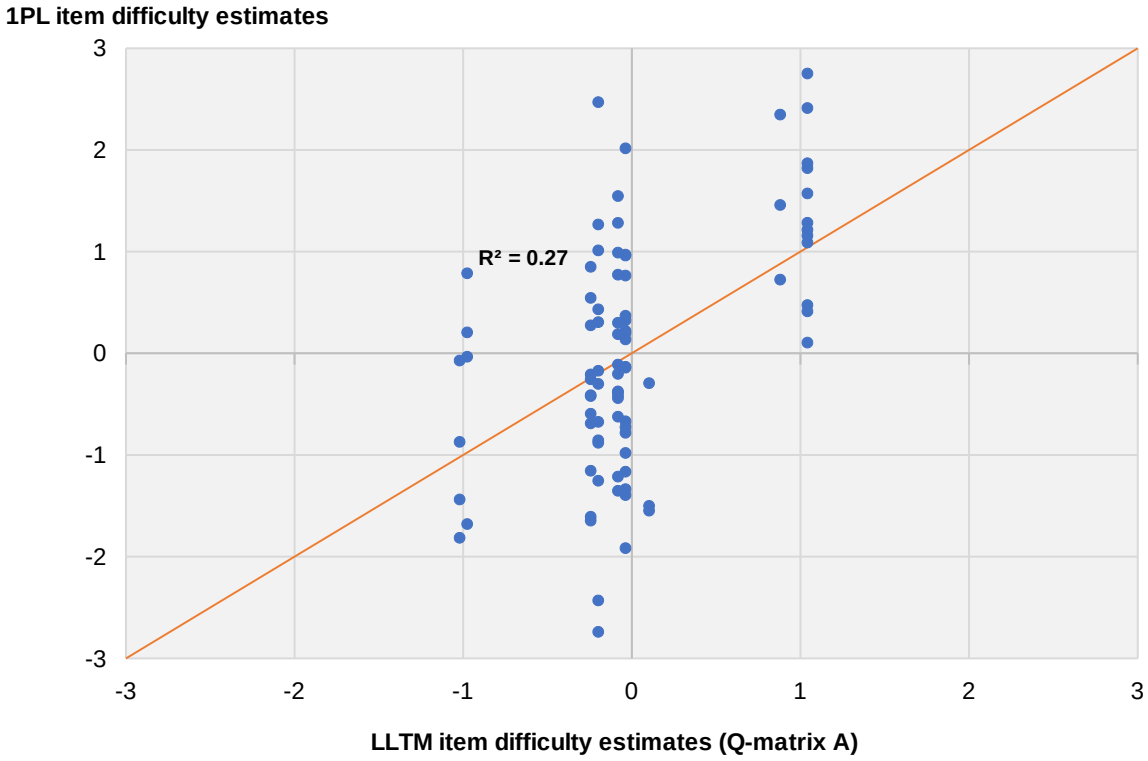
Table 19: Eta difficulty estimates for attributes in Q-matrix A

Item attribute	Eta parameter	*SE*	95% Lower CI	95% Upper CI	*p*
j_{A01} – Knowing	0	-	-	-	-
j_{A02} – Applying	0.045	0.008	0.028	0.061	<.001***
j_{A03} – Reasoning	1.124	0.010	1.103	1.144	<.001***
j_{A04} – Number	0	-	-	-	-
j_{A05} – Geometry	-0.162	0.008	-0.178	-0.146	<.001***
j_{A06} – Data	-0.939	0.012	-0.963	-0.915	<.001***

Eta parameters not calculated for two attributes due to Q-matrix structure.
CI = confidence interval. ***** = sig. at 95% confidence level, ****** = sig at 99% confidence, ******* = sig. at 99.9% confidence

Each item was associated with two attributes; one cognitive attribute, and one content attribute, and so the estimated difficulty of each item was the sum of the associated cognitive and content attribute. A normalisation constant based on the mean difficulty of the 85 items ($mean\,\beta = 0.082$) was also subtracted from each β_i estimate, such that the mean difficulty of the items as estimated under the LLTM was also equal to 0. The patterns of attributes create nine potential combinations of cognitive and content attributes, and accordingly, nine different item-level difficulty estimates. The overall correlation between the 1PL and LLTM item-level difficulty estimates was only moderate ¿), indicating that Q-matrix A was poor at accounting for the majority of variance in 1PL difficulty. The correlation between these estimates is represented in Figure 14, demonstrating many item difficulty estimates were poorly calibrated with the 1PL estimates (large deviance from the line of $y = x$).

Figure 14: Scatterplot of 1PL and LLTM item difficulty estimates (Q-matrix A)



While a linear regression approach shares many similarities to the LLTM, and produces similar item-level estimates and correlations to 1PL estimates, the regression approach on to 1PL

estimates produces vastly greater standard errors in item-level estimates. This is because regressing attributes on to 1PL estimates essentially treats the number of items as the sample size. Meanwhile, the LLTM has a sample size equal to the number of examinees and calculates directly from the item-level performance data. However, Green and Smith (1987) argued for the use of linear regression in addition to the LLTM, as a linear regression is the preferred method for exploring collinearity between attributes.

Tolerance statistics were calculated for each attribute, reflecting the percent of variance in a predictor attribute not accounted for by the other predictors. Following the approach employed by Chen et al. (2011), a tolerance value of .1 or lower was used as the cut-off value to identify attributes that may be redundant and better covered by other attribute(s). However, all four attributes ($j_{A01}$and j_{A04} not included) all had tolerance values greater than .75, suggesting very low collinearity. This is arguably to be expected given that, by the Q-matrix design, items could only belong to a single cognitive and content attribute.

In order to assess relative model fit between the 1PL and LLTM, Baghaei & Kubinger (2015) suggest a comparison using likelihood ratio tests. The difference between -2log-likelihoods of the 1PL and LLTM estimates are approximately chi-square distributed, with the difference in the number of parameters between the two models reflecting the number of degrees of freedom. At the 95% confidence level, the difference in -2log-likelihoods of the two models was 57431.35, far greater than the associated critical value of 101.88 of a chi-squared distribution with 80 degrees of freedom. The null hypothesis of no difference in the fit of the two models was rejected, and consequently, the 1PL was determined to have greater fit to the data than the LLTM under Q-matrix A. Q-matrix A therefore does not adequately account for the item-level attributes underlying the difficulty of these test-items.

5.3.2.2 Q-matrix B

This process was repeated for the 13 attributes in Q-matrix B. Q-matrix B consisted of the same three cognitive attributes and nine content attributes; j_{B01} (knowing) and j_{B05} (computations with whole numbers) were selected as the reference categories. There was an additional attribute for

whether the item required a constructed response. The eta parameters associated with each attribute are reported in Table 20. Once again, all included attributes had eta parameters that were significantly different to 0.

Table 20: Eta difficulty estimates for attributes in Q-matrix B

Item attribute	Eta parameter	*SE*	95% Lower CI	95% Upper CI	*p*
j_{B01} – Knowing	0	-	-	-	-
j_{B02} – Applying	-0.103	0.009	-0.122	-0.085	<.001***
j_{B03} – Reasoning	0.933	0.013	0.908	0.957	<.001***
j_{B04} – Number sense / place value awareness	-0.460	0.019	-0.498	-0.422	<.001***
j_{B05} – Computations with whole numbers	0	-	-	-	-
j_{B06} – Problem-solving with whole numbers	0.376	0.019	0.339	0.415	<.001***
j_{B07} – Understanding of equivalent fractions	-0.464	0.023	-0.509	-0.419	<.001***
j_{B08} – Computations with fractions and decimals	0.149	0.022	0.105	0.193	<.001***
j_{B09} – Patterns and relationships	-0.521	0.021	-0.562	-0.480	<.001***
j_{B10} – Points, lines and angles	-0.465	0.019	-0.503	-0.427	<.001***
j_{B11}– 2D and 3D Shapes	-0.114	0.017	-0.147	-0.080	<.001***
j_{B12} – Data	-0.934	0.019	-0.971	-0.897	<.001***
j_{B13} – Item requires a constructed response	0.237	0.008	0.222	0.253	<.001***

Eta parameters not calculated for two attributes due to Q-matrix structure.
CI = confidence interval. * = sig. at 95% confidence level, ** = sig at 99% confidence, *** = sig. at 99.9% confidence

The correlation between 1PL estimates and LLTM estimates under Q-matrix showed only marginal improvement from Q-matrix A ($r = .58, r^2 = .34$). Comparing model fit showed that once

again, the 1PL was the far greater fitting model (difference in -2log-likelihoods of 50676.1 compared to a critical value of 93.9). The tolerance value for constructed-response items (j_{B13}) was highest, at .88. Tolerance values for all attributes ranged between .23 and .62 for all other attributes, suggesting low collinearity between attributes.

5.3.2.3 Q-matrix C

The eta-parameters of the 15 attributes in Q-matrix C are reported in Table 21. Q-matrix C added two reading-related attributes to Q-matrix B, both of which produced eta parameters that were significantly higher than 0; this means that items that were judged by the IEA to have high reading demands, or those that contained at least one use of mathematical or technical language had a statistically significant effect on item difficulty estimates. However, the actual impacts of these on the overall difficulty estimates were both small.

Table 21: Eta difficulty estimates for attributes in Q-matrix C

Item attribute	Eta parameter	*SE*	95% Lower CI	95% Upper CI	*p*
j_{C01} – Knowing	0	-	-	-	-
j_{C02} – Applying	-0.106	0.009	-0.125	-0.087	<.001***
j_{C03} – Reasoning	0.871	0.013	0.845	0.896	<.001***
j_{C04} – Number sense / place value awareness	-0.486	0.019	-0.524	-0.448	<.001***
j_{C05} – Computations with whole numbers	0	-	-	-	-
j_{C06} – Problem-solving with whole numbers	0.315	0.020	0.277	0.354	<.001***
j_{C07} – Understanding of equivalent fractions	-0.490	0.023	-0.535	-0.444	<.001***
j_{C08} – Computations with fractions and decimals	0.166	0.022	0.122	0.211	<.001***
j_{C09} – Patterns and relationships	-0.621	0.021	-0.663	-0.580	<.001***
j_{C10} – Points, lines and angles	-0.554	0.020	-0.592	-0.515	<.001***
j_{C11} – 2D and 3D shapes	-0.177	0.017	-0.211	-0.142	<.001***
j_{C12} – Data	-1.099	0.021	-1.139	-1.059	<.001***
j_{C13} – Item requires a constructed response	0.216	0.008	0.200	0.231	<.001***
j_{C14} – High reading demands	0.220	0.010	0.201	0.240	<.001***
j_{C15} – Contains mathematical / technical language	0.051	0.008	0.035	0.067	<.001***

Eta parameters not calculated for two attributes due to Q-matrix structure.
CI = confidence interval. * = sig. at 95% confidence level, ** = sig at 99% confidence, *** = sig. at 99.9% confidence

The addition of these two additional attributes had negligible impact on the correlation between 1PL and LLTM difficulty estimates ($r = .59, r^2 = .34$), and similarly poor model fit relative to the 1PL (difference in -2log-likelihoods of 50155.1 compared to a critical value of 91.7) compared to Q-

matrix B. Collinearity tolerance values ranged between .22 and .62 for the cognitive and content attributes shared with Q-matrix B, while the tolerance values for j_{C13} (high reading demand) and j_{C14} (mathematical / technical language) were both higher (.63 and .87 respectively), indicating that there was very low collinearity between the reading-related attributes and the cognitive and mathematical content attributes.

5.3.2.4 Q-matrix D

Following the use of the three Q-matrices designed around the IEA's assessment framework, analysis moved on to the use of the 19-attribute framework adapted from Tatsuoka et al. (2004). Unlike Q-matrices A-C, Q-matrix D needed to be applied retroactively to each of the 85 items. All 85 items were independently coded, with 30 of these items being selected randomly for reliability checks by two other coders. Across these 30 items, disagreement only arose regarding six individual item-level codes. After discussion, all but one of these remained as they were previously, while one item's codes were changed in light of these discussions.

Each item belonged to one of the five content attributes, with j_{D01} (whole numbers) being selected as a reference category. The structure of Q-matrix D is otherwise less rigid than Q-matrices A-C, with a much greater variety of possible attribute combinations. The associated eta parameters as estimated under the LLTM for Q-matrix D are reported in Table 22. All but two of the attribute eta parameters (j_{D18} and j_{D19}) were found to differ significantly from 0, but as with previous Q-matrices, several attributes had statistically significant eta parameters that only had small impacts on item difficulty estimates.

Table 22: Eta difficulty estimates for attributes in Q-matrix D

Item attribute	Eta parameter	*SE*	95% Lower CI	95% Upper CI	*p*
j_{D01} – Whole numbers	0	-	-	-	-
j_{D02} – Fractions and decimals	0.081	0.015	0.052	0.111	<.001***
j_{D03} – Geometry	-0.049	0.013	-0.075	-0.022	<.001***
j_{D04} – Data	-0.866	0.020	-0.905	-0.827	<.001***
j_{D05} – Measuring or estimating in real-world contexts	0.721	0.017	0.688	0.755	<.001***
j_{D06} – Translate/formulate equations and expressions	0.083	0.012	0.058	0.108	<.001***
j_{D07} – Perform computations in arithmetic / geometry	0.826	0.011	0.805	0.847	<.001***
j_{D08} – Make judgements or evaluate correctness	0.420	0.011	0.398	0.442	<.001***
j_{D09} – Logical reasoning in written contexts	0.926	0.013	0.899	0.952	<.001***
j_{D10} – Logical reasoning in visual contexts	0.725	0.014	0.697	0.753	<.001***
j_{D11} – Number-sense / place-value awareness	0.293	0.014	0.266	0.321	<.001***
j_{D12} – Patterns and sequences	-0.212	0.016	-0.242	-0.181	<.001***
j_{D13} – Proportional relationships	1.262	0.024	1.215	1.309	<.001***
j_{D14} – Compare entities	0.214	0.012	0.191	0.237	<.001***
j_{D15} – Use visual information	0.056	0.012	0.031	0.080	<.001***
j_{D16} – Check multiple-choice options	0.248	0.014	0.220	0.276	<.001***
j_{D17} – Open-ended response	0.667	0.011	0.646	0.688	<.001***
j_{D18}– Question requires verbal interpretation	-0.008	0.010	-0.029	0.012	.413
j_{D19} – Novel / non-routine contexts	-0.018	0.010	-0.037	0.002	.074

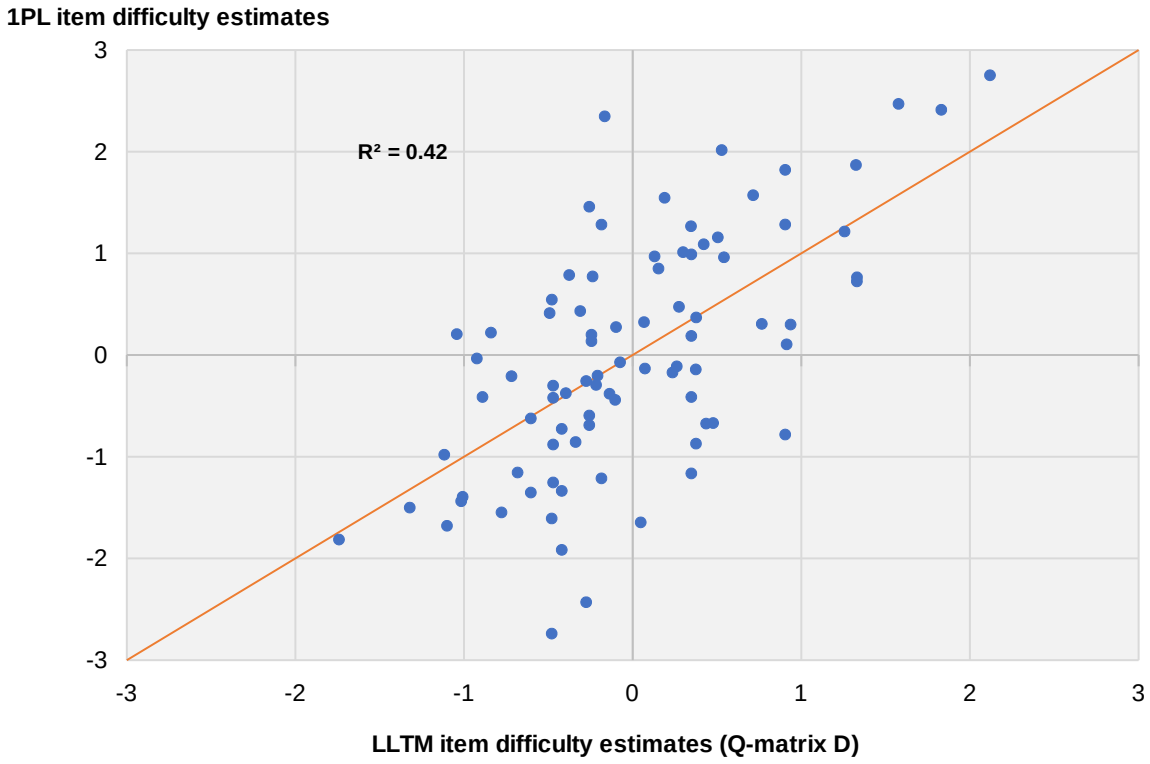
Eta parameters not calculated for one attribute due to Q-matrix structure.
CI = confidence interval. * = sig. at 95% confidence level, ** = sig at 99% confidence, *** = sig. at 99.9% confidence

The mean difficulty of items under the LLTM was comparatively high under Q-matrix D ($mean\,\beta = 1.145$). This was to be expected given that, under Q-matrix D, items had an average of five attributes, whereas all items had just two attributes in matrices A and B, and it would generally be expected that items involving more content or skill attributes would be more difficult. This is reflected by most attributes in Q-matrix D having positive eta parameters. The mean difficulty was subtracted from each item as a normalisation constant to equate the mean to 0 to facilitate comparisons to the 1PL estimates.

Items that required students to perform arithmetic or geometric computations, that required students to reason logically about either verbally, or visually presented information, and problems relating to measurement or estimation contexts were among those that had the largest positive effects on difficulty. Open-ended response items were associated with around a 0.67 logit increase in item difficulty relative to multiple-choice items. The only attribute associated with a large decrease in item difficulty estimates was whether the item tested the data cognitive domain, consistent with the findings from Q-matrices A-C.

The correlation between 1PL and LLTM estimates exhibited some improvement over Q-matrices A-C ($r = .65, r^2 = .42$), but still explained less than half of the variance in 1PL item difficulty estimates. Figure 15 presents the comparison of item difficulty estimates under the 1PL and LLTM for Q-matrix D.

Figure 15: Scatterplot of 1PL and LLTM item difficulty estimates (Q-matrix D)



Tolerance statistics did not indicate any major issues of collinearity, with tolerance values ranging between .34 and .73. However, the 1PL was still found to have significantly better fit to the data than the LLTM (difference in -2log-likelihoods of 46036.0 compared to a critical value of 86.0).

5.3.2.5 Relative fit of Q-matrices C and D

A final comparison was made between the relative fit of Q-matrices C and D, to compare model fit between the best fitting of the three IEA frameworks and that derived from Tatsuoka et al. (2004). Q-matrix C was compared in reference to Q-matrix D. At the 95% confidence level, the difference in -2log-likelihoods of the two models was 4838.6, far greater than the associated critical value of 12.6. The null hypothesis of no difference in the fit of the two models was rejected, and consequently, Q-matrix D was determined to have greater fit to the data than Q-matrix C.

5.4 Discussion

This paper looked to investigate the item-level attributes that contribute to the difficulty of items in the fourth-grade mathematics assessments of TIMSS 2011. The analysis focused on a cohort of just over 20,000 students across six Anglophone countries and their responses to 85 different test-items, and primarily explored whether the mathematics assessment framework, as defined by the IEA, is predictive of item difficulty. Comparisons were made to an alternative item framework first used by Tatsuoka et al. (2004) in their assessment of skill mastery in mathematics in TIMSS at the eighth grade.

The results of the 1PL model show that mean ability levels were slightly higher than the mean of item difficulties, indicating that these test-items were somewhat on the easy side for this cohort of students. There was however a good range of item difficulties across the range of examinee abilities. A small number of items were only well targeted to very small numbers of low ability students, while there were a greater number of students for whom all the items were too easy. It is important to remember of course that TIMSS has much greater international reach than these six countries, with five of these having average scores that were significantly above the TIMSS scale centerpoint score of 500; the sub-cohort therefore reflects a higher-performing population, and so it is perhaps unsurprising that the mean ability of this group was higher than the mean estimated item difficulty. Overall, the difficulty level of the TIMSS fourth-grade mathematics assessment appeared to be well targeted to the ability levels of this sub-cohort of students.

Four frameworks for exploring item difficulty were investigated. The first three of these were derived from item-level categorisations used by the IEA. The first focused on the broad level cognitive and content domains used in TIMSS 2011, while the second provided greater specificity on the content of these items as outlined in the TIMSS 2011 assessment framework. The third framework incorporated two item-level readability metrics used by the IEA in their TIMSSPIRLS relationships report (Martin & Mullis, 2013). Across all three of these frameworks, the correlations between 1PL and LLTM item difficulty estimates were somewhat low, explaining between 27% and 34% of the variance in item difficulties. Other item-level attributes not accounted for in these

frameworks therefore account for most of the variance in item difficulty. When assessing model fit, all three of these frameworks showed significantly worse fit to the data than the 1PL. This suggests that the item-level attributes, as defined in the TIMSS 2011 assessment framework and assigned to individual items, do not explain why some items were more difficult for these students than others. It should of course be noted that the IEA do not claim that item difficulties in TIMSS are driven solely by these factors. However, these findings do suggest that additional caution be taken when interpreting content domain and cognitive domain subscale results.

Though Q-matrix D, adapted from the Q-matrix employed by Tatsuoka et al. (2004), also had weaker fit to the data than the 1PL, it had significantly better fit to the data than Q-matrix C, and was more strongly correlated with the 1PL's difficulty estimates, explaining around 42% of the variance in item difficulties. This is substantially higher than often reported in item difficulty modelling studies which typically explain less than 20% of this variance (El Masri et al., 2017). Embretson and Daniel (2008) argue that when retrofitting to existing assessments, an LLTM approach should be considered strong if it can explain around half of the variance in item difficulties. While this threshold has not been reached, the application of Q-matrix D to the TIMSS 2011 dataset should be seen as having produced a promising result. Further work in this area is needed. However, the relatively stronger fit of the alternative framework does raise questions about the validity of other cognitive modelling studies that have relied on the TIMSS assessment framework for defining the attributes under investigation.

One surprising finding was that the introduction of reading-related attributes in Q-matrix C had almost no impact on the proportion of explained variance relative to Q-matrix B. Studies that have focused on either mathematics assessments generally (e.g. Abedi & Lord, 2001; Rewer & Greefrath, 2023) or in TIMSS science (e.g. Rosca, 2004) which used similarly structured items and assessment frameworks, have typically reported significant effects of reading-related factors on item difficulty. However, a similar result was found for the readability attribute in Q-matrix D, with items requiring higher-level verbal interpretation not significantly affecting the estimate of the item's difficulty. One interpretation is that the reading demands in TIMSS mathematics test-items

are low and well-controlled. Another explanation is that the reading demands as experienced by students are not well captured by either of these Q-matrices.

A few limitations of the presented analyses should also be discussed. One limitation arises from the trade-off between the breadth and depth of attribute coverage. Q-matrices B and C, for example, do not represent the full depth of the TIMSS 2011 mathematics assessment framework because, even using 85 items, there is not sufficient coverage of the topic bullets across the items to estimate more specific eta parameters reliably. This is particularly the case for items belonging to the geometry and data content domains. Further, there is no expansion on the cognitive domains; while the TIMSS 2011 assessment framework does provide a greater summary of the skills comprising each of the cognitive domains, these are not designated to individual items. Greater analysis of how these are represented and tested by these items might allow for more nuanced discussion of the impact of cognitive skills on item difficulty estimates and improve model fit.

Similar issues applied to the retrofitting of Q-matrix D; Tatsuoka et al.'s (2004) framework was originally designed for items targeting the eighth-grade TIMSS assessment, and so some attributes simply did not apply to the fourth-grade dataset. However, other relevant attributes had to be removed or amended because there was insufficient coverage of these attributes across the 85 items. It is therefore possible that this analysis has underestimated the extent to which the TIMSS assessment framework (and/or Tatsuoka et al.'s) framework could explain item difficulty because of data limitations.

The structure of the TIMSS framework and how it is applied to items also introduced challenges – with each item belonging to exactly one cognitive attribute and content attribute in Q-matrices A-C, there was a less-than-ideal structure to the Q-matrix that may not have best represented the range of content and skills tested by individual items. This was somewhat mitigated in Q-matrix D, which had more complex patterns of attributes. Another broader potential issue relates to the use of the LLTM, and its assumption that the contributions of different attributes associated with item difficulty stack linearly. In practice, there may be more complex interactions between

attributes that the LLTM does not account for. The LLTM may therefore underestimate the strength of the relationship between the selected attributes and item difficulties.

In summary, the results presented here raise questions about how well the TIMSS mathematics assessment framework relates to the TIMSS mathematics items themselves. If, as suggested here, the assessment framework is poorly linked to the difficulty of the TIMSS mathematics test-items, then healthy caution should be applied when interpreting more specific claims about students' performance in TIMSS mathematics, such as reporting of scores at the subscale level. This also has implications on the validity of the growing body of literature that has made use of the TIMSS mathematics assessment framework for cognitive diagnostic purposes. While the IEA does not claim that the assessment framework explains or accounts for all the sources of difficulty in TIMSS items, this has arguably been an implicit assumption of much of this research. Analyses focused on item-level performance in TIMSS mathematics should be confident that their cognitive models for describing and explaining student performance are strongly connected to the difficulties of the test-items. This is not an assumption supported by the analysis presented here.

5.5 References

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14(3*), 219-234. https://doi.org/10.1207/S15324818AME1403_2

Baghaei, P., & Kubinger, K.D. (2015). Linear logistic test modeling with R. *Practical Assessment, Research & Evaluation, 20(1).* https://doi.org/10.7275/8f33-hz58

Birenbaum, M., Tatsuoka, C., & Xin, T. (2005). Large-scale diagnostic assessment: comparison of eighth graders' mathematics performance in the United States, Singapore and Israel. *Assessment in Education: Principles, Policy & Practice, 12(2)*, 167-181. https://doi.org/10.1080/09695940500143852

Chen, Y.H. (2011). Cognitive diagnosis of mathematics performance between rural and urban students in Taiwan. *Assessment in Education: Principles, Policy & Practice, 19(2)*, 193-209. https://doi.org/10.1080/0969594X.2011.560562

Chen, Y., MacDonald, G., & Leu, Y. (2011). Validating cognitive sources of mathematics item difficulty: Application of the LLTM to fraction conceptual items. *The International Journal of Educational and Psychological Assessment, 7(2*), 74-93. Retrieved from: https://www.researchgate.net/profile/Yi-Hsin-Chen-3/publication/286926672_Validating_cognitive_sources_of_mathematics_item_difficulty_Application_of_the_LLTM_to_fraction_conceptual_items/links/567029e408ae5252e6f1d43a/Validating-cognitive-sources-of-mathematics-item-difficulty-Application-of-the-LLTM-to-fraction-conceptual-items.pdf

Delafontaine, J., Chen, C., Park, J. Y., & Van den Noortgate, W. (2022). Using country-specific Q-matrices for cognitive diagnostic assessments with international large-scale data. *Large-scale Assessments in Education, 10(1)*, 19, https://doi.org/10.1186/s40536-022-00138-4

El Masri, Y.H., Ferrara, S., Foltz, P.W. and Baird, J. (2017). Predicting item difficulty of science national curriculum tests: the case of key stage 2 assessments. *The Curriculum Journal, 28*, 59-82. https://doi.org/10.1080/09585176.2016.1232201

Fischer, G.H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*, 359-374. https://doi.org/10.1016/0001-6918(73)90003-6

Green, K.E., & Smith, R.M. (1987). A comparison of two methods of decomposing item difficulties. *Journal of Educational Statistics, 12*, 369-381. https://doi.org/10.2307/1165055

Hambleton, R.K., & Jirka, S.J. (2006). Anchor-based methods for judgementally estimating item statistics. In: S.M. Downing & T.M. Haladyna (Eds.), *Handbook of test development* (pp. 399–420). Mahwah, NJ: Lawrence Erlbaum.

Holmes, S.D., Meadows, M., Stockford, I., & He, Q. (2018). Investigating the comparability of examination difficulty using comparative judgement and Rasch modelling. *International Journal of Testing, 18(4)*, 366–391. https://doi.org/10.1080/15305058.2018.1486316

Lee, Y.S., Park, Y.S., & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the U.S. national sample using the TIMSS 2007. *International Journal of Testing, 11(2)*, 144-177. https://doi.org/10.1080/15305058.2010.534571

Mair, P. & Hatzinger, R. (2007). Extended Rasch modelling. The eRm package for the application of IRT models in R. *Journal of Statistical Software, 20*, 1-20. https://doi.org/10.18637/jss.v020.i09

Martin, M.O. & Mullis, I.V.S. (2012). *Methods and procedures in TIMSS and PIRLS 2011.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from: https://timssandpirls.bc.edu/methods/index.html

Martin, M.O. & Mullis, I.V.S. (2013). *TIMSS and PIRLS 2011: Relationships among reading, mathematics, and science achievement at the fourth grade - implications for early learning.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from: https://timssandpirls.bc.edu/timsspirls2011/international-database.html

Melican, G.J., Mills, C.N., & Plake, B.S. (1989). Accuracy of item performance predictions based on the Nedelsky standard setting method. *Educational and Psychological Measurement, 49,* 467-478. https://doi.org/10.1177/0013164489492020

Messick, S. (1993). Validity. In: R.L. Linn (Ed.), *Educational measurement, 3*, 3-103. New York, NY: Macmillan.

Mullis, I.V.S., Martin, M.O., Ruddock, G.J., O'Sullivan, C.Y., Arora, A., & Erberber, E. (2005).*TIMSS 2007 assessment frameworks*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from: https://timssandpirls.bc.edu/timss2007/frameworks.html

Park, Y.S., & Lee, Y.S. (2014). An extension of the DINA model using covariates: examining factors affecting response probability and latent classification. *Applied Psychological Measurement, 38(5)*, 376-390. https://doi.org/10.1177/0146621614523830

Pettersen, A., Nortvedt, G.A. (2018). Identifying competency demands in mathematical tasks: recognising what matters. *International Journal of Science and Mathematics Education 16*, 949-965. https://doi.org/10.1007/s10763-017-9807-5

Pollitt, A., Ahmed, A., & Crisp, V. (2007). The demands on examination syllabuses and question papers. In: P. Newton, J.A. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards*, 166-206. London: Qualifications and Curriculum Authority.

Rewer, A., & Greefrath, G. (2023). The impact of linguistic and mathematical difficulty-determining characteristics of students' performance in test-items in mathematics. In: *Thirteenth Congress of the European Society for Research in Mathematics Education (CERME13) (No. 21)*. Alfréd Rényi Institute of Mathematics; ERME. Retrieved from: https://hal.science/hal-04413560v1

Rosca, C.V. (2004). *What makes a science item difficult? A study of TIMSS-R items using regression and the linear logistic test model*. Doctoral dissertation, Boston College. Retrieved from: https://www.proquest.com/dissertations-theses/what-makes-science-item-difficult-study-timss-r/docview/305214920/se-2

Tatsuoka, K K., Corter, J.E., & Tatsuoka, C. (2004). Patterns of diagnosed mathematical content and process skills in TIMSS-R across a sample of 20 countries. *American Educational Research Journal, 41(4)*, 901-926. https://doi.org/10.3102/00028312041004901

Toker, T., & Green, K. (2012). *An application of cognitive diagnostic assessment on TIMMS-2007 8th grade mathematics items*. Paper presented at the annual meeting of the American Educational Research Association (Vancouver, British Columbia, Canada, April 14-16, 2012). Retrieved from: https://eric.ed.gov/?id=ED543803

Yamaguchi, K., & Okada, K. (2018). Comparison among cognitive diagnostic models for the TIMSS 2007 fourth grade mathematics assessment. *PloS One, 13(2),* https://doi.org/10.1371/journal.pone.0188691

6 Paper 3: Modelling students' content knowledge and skill mastery in mathematics: an approach using data from TIMSS

The datasets from international large-scale assessment studies such as TIMSS, PISA and PIRLS have been increasingly used in cognitive diagnostic modelling analyses. Concerns have been raised more broadly about retrofitting of diagnostic models to assessment datasets not designed with cognitive diagnoses in mind, but there has been little discussion of the issues of retrofitting specifically in the context of international assessments. This paper discusses these concerns, providing demonstrations of how these apply to international large-scale assessment data. In this paper, the DINA model, one of the foundational cognitive diagnosis models, was applied to item-level data from the fourth-grade mathematics assessment of TIMSS 2011. The results of these analyses demonstrate how and why the application of cognitive diagnostic models to international large-scale assessment datasets are limited and should be interpreted with caution and scepticism.

6.1 Introduction

In recent decades, the results from international large-scale assessments such as the Trends in International Mathematics and Science Study (TIMSS), Programme for International Student Assessment (PISA) and Progress in International Reading Literacy Study (PIRLS) have become increasingly influential in the evaluation of global educational standards and in informing national educational policies (Addey, 2017; Gorur, 2017; Fischman et al., 2018). However, the conclusions drawn from these studies based on score differences between countries, or groups of students within those countries are only descriptive – they may suggest that students in group A have higher levels of mathematics performance than students in group B, but they do not tell us about the more specific strengths or weaknesses in students' mathematical abilities that contribute to those score differences (Wagemaker, 2020). Further, it has been argued that the

results of large-scale international assessments fail to provide useful information that can be directly applied at the classroom level (e.g. Lee et al., 2011; Holliday & Holliday, 2003; Wang, 2001; Long, 2007). Consequently, there has been a drive to expand upon what is reported when the results of these studies are published through secondary analyses of their publicly available datasets.

Meanwhile, recent decades have witnessed the growing number and popularity of statistical models that extract highly granular diagnostic information about student's specific strengths and weaknesses from item-level performance data. In particular, there has been a surge of interest in a number of different cognitive diagnostic models (CDMs). CDMs are latent class models that can be used to make finer-grained diagnostic claims about the cognitive skills and knowledge states that are measured by individual test-items, and make inferences about how these are reflected by examinees' item-level performances (Leighton & Gierl, 2007; Rupp, 2007; Williamson, 2023; Zhang, 2023). Cognitive skills and knowledge states as measured by the items and 'mastered' by examinees are commonly referred to as attributes.

CDMs have the potential to provide strong, and highly specific inferences about students' individual current strengths, weaknesses and areas for development, and could therefore act as powerful diagnostic tools to inform further teaching and learning. However, Ravand and Robitzsch (2015) lamented the relative lack of real-world applications of CDMs, noting that CDMs typically require larger sample sizes than most smaller-scale assessments and pilot studies can produce. As such, there has been limited application of these models beyond simulation studies.

On the surface, the datasets provided by international large-scale assessment studies such as TIMSS, PISA and PIRLS appear well suited to these kinds of diagnostic modelling. These studies offer interested researchers with free access to item-level performance datasets with very large nationally representative samples of students, and their test-items have already undergone thorough piloting to ensure they are well targeted to the ranges of ability within their assessment populations. Given the wider interest in extracting performance information beyond that provided by the unidimensional item-response theory (IRT) approaches used in the scoring and reporting

of international assessment studies, it is perhaps unsurprising that many published research articles have already sought to capitalise on the opportunities seemingly afforded by international assessment data. Stiff (see paper 1, Chapter 4) noted that, of these studies and subjects, item-level data from TIMSS mathematics been the most frequently used in cognitive diagnostic modelling analyses, particularly over the last fifteen years.

Though the datasets of these international large-scale assessments may appear attractive for researchers looking to expand upon international assessment results, or those interested more broadly in exploring cognitive diagnostic modelling, the use of CDMs to these data should also draw concern. Perhaps the central argument against the use of CDMs to analyse international assessment data is simply that CDMs are best applied to assessment datasets designed specifically for the purpose of providing item-level diagnostic information (Gierl & Cui, 2008; Ravand & Baghaei, 2020; Williamson, 2023). This is not the case with the assessments included in TIMSS, PISA, or PIRLS, which are specifically designed to provide valid information a) at the national level, rather than at the level of individual students, and b) about broad subject domain performance (e.g. mathematics, science), rather than for very granular subtopics within those subjects. Deonovic et al. (2019) argued that retrofitting CDM analyses are complex “when the breadth of the domain is relatively wide” (p. 446), as is the case with these international assessments. These concerns apply more widely than just to international assessment data. Ravand and Baghaei (2020) observed that applications of CDMs to assessment datasets not originally designed with cognitive diagnosis in mind outnumber ‘true’ applications of CDMs by around four to one, but argued that these ‘retrofitting’ analyses should be considered a ‘measure of last resort’ (p.27) in the use of CDMs. However, given the rising popularity of CDM applications to international assessment data (see paper 1, Chapter 4), the wider concerns about the applications of CDMs have seemingly not deterred researchers in applying those models to these datasets.

This paper looks to discuss and demonstrate the limitations and issues associated specifically with the application of CDMs to international assessment data. The focus here is not on a critique of either CDMs or international assessment data individually, but rather on the specific practice of

applying CDMs to international assessment data. Practical issues in the application of CDMs and the validity of the results produced by those analyses will be discussed with reference to example applications of the DINA model, one of the most widely used and easily interpretable CDMs, to data from TIMSS mathematics.

6.2 Literature review

6.2.1 Cognitive diagnostic models and Q-matrices

Though the varying CDMs differ in their underlying assumptions about how examinees' cognitive skills and knowledge states are reflected in performance at a test or item level, almost all CDMs share similar approaches in classifying examinees into groups of mastery and/or non-mastery of attributes. CDMs express examinees' performance levels in terms of probabilities of attribute mastery and non-mastery. This can either be in terms of mastery or non-mastery of single attributes, or probabilities of belonging to each latent class of masteries (Lee & Sawaki, 2009).

Each examinee, e, is characterised by a binary attribute vector $a_e=(a_{e1}\ldots a_{eJ})$, where $a_{ej}=1$ represents the examinee's mastery of attribute j, and $a_{ej}=0$ represents non-mastery.

Most CDMs make use of a Q-matrix (Tatsuoka, 1985), which can be described as a binary matrix of i item rows and j attribute columns that specifies which attributes are measured by each item. If item i is believed to measure attribute j, this is marked at intersection Q_{ij} in the Q-matrix with a 1, or else is filled with a 0. A small example of how a Q-matrix might be constructed for a test of 15 items is presented in Table 23.

Table 23: Example of a Q-matrix

Item	Attribute j_1	Attribute j_2	Attribute j_3	Attribute j_4	Attribute j_5	Attribute j_6
i_1	1	0	0	0	0	0
i_2	0	1	0	0	0	0
i_3	1	1	0	0	0	0
i_4	0	0	1	0	0	0
i_5	0	1	1	0	0	0
i_6	0	0	0	1	0	0
i_7	1	0	0	1	0	0
i_8	0	0	0	0	1	0
i_9	0	0	0	1	1	0
i_{10}	0	0	0	0	0	1
i_{11}	1	0	0	0	0	1
i_{12}	0	0	1	0	1	0
i_{13}	0	1	0	1	0	0
i_{14}	1	0	1	0	0	1
i_{15}	0	0	1	0	1	1

The development and validation of the most suitable Q-matrix is arguably the most critical step in conducting cognitive diagnostic analysis (Gorin, 2009). Gu and Xu (2021) argued that there are three minimal conditions that a Q-matrix must fulfil for it to be considered ‘identifiable’. The identifiability of a Q-matrix is how well it can provide unique classifications of examinees into their true profiles of attribute mastery and non-mastery. It is not a necessary pre-requisite for conducting cognitive diagnostic analysis, but rather, they argued that Q-matrices that do not fulfil these conditions run a strong risk of misspecification. These three conditions are:

- **Completeness** – each attribute of the Q-matrix must be measured in isolation by at least one test-item. This assists in isolating the effects of mastery or non-mastery of each specific attribute, as it can be difficult to isolate individual masteries when the interactions of attributes are complex for all items.
- **Distinctness** – each attribute of the Q-matrix must represent a unique skill or knowledge state; they can be related, but it must be possible for an examinee to possess mastery of one attribute but not the other.

- **Repetition** – examinees should have at least three responses for each attribute of the Q-matrix. This helps to account for the effects of item-level noise (discussed further in section 6.2.2). Others (e.g. Jang, 2009) have suggested three items per attribute is still too low, but this can be difficult to achieve if Q-matrices involve more than five attributes.

Additionally, assessments designed with cognitive diagnostic modelling in mind will typically consist of a mix of both items that measure individual attributes in isolation (meeting the condition of completeness) as well as items that assess multiple attributes simultaneously (assisting in meeting the condition of repetition). The simultaneous testing of multiple attributes through a single item sometimes referred to as having 'within-item multidimensionality' (Baghaei, 2012) and should ideally complement the test-items that isolate the measurement of single attributes. Overall, the collective structure of items used in an assessment should ensure sufficient coverage of each attribute for reliable estimations of attribute mastery at the examinee-level.

However, ensuring the validity of the Q-matrices used in CDMs is often challenging. One of the greatest threats to the validity of CDM analyses is misspecification of the Q-matrix. There are many potential causes of such misspecification, depending on how the Q-matrix was constructed. In most applications of CDMs, Q-matrices are constructed by subject-experts and/or those involved in test-item development. There can often be subjectivity and bias in how these experts perceive items and how they anticipate examinees to interact with them, meaning that multiple experts may produce substantially different Q-matrices for the same sets of items and attributes. These issues can be exacerbated if there is ambiguity in how the attributes themselves are defined, or if there are redundant attributes that share too many similarities with other attributes of the Q-matrix (i.e. the Q-matrix does not meet the condition of distinctness).

6.2.2 The DINA model

Of the many CDMs in use, perhaps the most parsimonious is the deterministic inputs, noisy 'and' gate (DINA) model (Junker & Sijtsma, 2001; de la Torre & Douglas, 2004). The DINA model has been the foundational model for many other approaches to cognitive diagnosis and assessment (Junker & Sijtsma, 2001); while the DINA model itself has some very strict assumptions, it is

mathematically flexible and has allowed for the generation of more generalised models with looser rules for parameterisation (e.g. de la Torre, 2012), or for models incorporating other variates such as time for assessments conducted over multiple sessions (e.g. Ma et al., 2025).

The DINA model assumes a conjunctive (non-compensatory) relationship between attributes in which an examinee must have mastered all attributes required by an item to provide a correct response. For every interaction between examinee e and item i, the ideal response pattern (e.g. a correct response, denoted by n_{ei}) is calculated as the product of the mastered attributes for the item:

$$n_{ei}=\prod_{j=1}^{J}\alpha_{ej}^{Qij}$$

If an examinee has mastered all attributes required by an item, then $n_{ei}=1$. If however, an examinee has not mastered any one of the j attributes required by an item, $n_{ei}=0$.

The DINA model incorporates item-level slipping (s_i) and guessing (g_i) parameters to represent noise in examinee's responses. The naming of these parameters reflect the assumptions of the sources of noise in the DINA model. Slipping occurs when examinees fail to provide the correct response to a test-item that the DINA model estimates they have the required masteries of. Conversely, guessing occurs when examinees provide the correct response despite not having mastered all of the required attributes.

While s_i and g_i may reflect examinees' true experiences of failing to provide correct answers to questions despite sufficient knowledge or skill, or random selection/production of expected responses, high item-level s_i and g_i values are far more likely to represent underlying issues such as poor model fit, or Q-matrix misspecification (Junker & Sijtsma, 2001).

6.2.3 DINA and international assessments

Lee et al. (2011) were among the first to apply the DINA model to data from TIMSS. Their analysis looked to further expand on the reported score differences between the United States

(USA) main sample for TIMSS 2007, and those reported for Massachusetts and Minnesota, who entered TIMSS 2007 with additional benchmarking samples. Students in the Massachusetts and Minnesota samples had significantly higher average scores than the main USA cohort. Lee et al. (2011) applied a Q-matrix based on 15 content topic areas listed in the TIMSS 2007 mathematics assessment framework to 25 fourth-grade test-items to try and explain this difference in terms of attribute masteries. That is, they looked to investigate whether these overall score differences could be explained by any specific areas of strength within the Massachusetts and Minnesota samples relative to the USA cohort.

While the TIMSS mathematics frameworks used in each cycle of the study specify that each item tests a single content domain, and the attributes of the Q-matrix used by Lee et al. (2011) were derived from the 2007 version of this framework (Mullis et al., 2005), their Q-matrix allowed for multiple attributes across content domains. For example, geometry items could have attributes derived from topics within both the geometry and number content domains. Each item in their Q-matrix was categorised as having assessed between one and six of the attributes.

Although students in Massachusetts had the highest proportion of correct answers to these test-items, the average proportion of attribute mastery, as estimated by the DINA model, was lower than in the other two cohorts. In combination with the calculated guessing and slip parameters produced with the basic DINA model, it was suggested that students in Massachusetts scored well overall in TIMSS 2007 because they were more successfully able to guess the correct responses to test-items for which they had not mastered the required attributes, particularly in the ‘data display' content domain. Conversely, students in the USA sample had particularly large slipping parameters on four items; this included two items in the ‘number' content domain (with entirely different attributes), as well as two items in the ‘geometric shapes and measures' content domain both requiring the same attribute related to the calculation/estimation of perimeter, area and/or volume. Follow-up logistic regressions were used to compare the significance of attribute mastery differences between the three cohorts. Students in the USA sample had significantly higher mastery of two attributes, relating to calculation/estimation of perimeter area and/or volume, and reading data from tables/pictographs/bar charts/pie charts, though significantly lower

mastery of three attributes in the ‘number' content domain, and one in each of the other content domains.

Lee et al.’s (2011) use of the DINA to assess attribute mastery in TIMSS has been used as the basis for further research focusing on cognitive diagnostic modelling of TIMSS mathematics. For example, Hansen et al. (2014) followed up on Lee et al.’s work by suggesting minor changes to the Q-matrix to improve model fit, and to better account for local dependence between two items used in the analysis. Other studies meanwhile have used simplified Q-matrices with a smaller number of attributes based on the content domains and cognitive domains of the TIMSS mathematics frameworks (Hou et al., 2013; George & Robitzsch, 2018). More recently, Delafontaine et al. (2022) suggested that the Q-matrices used in DINA analyses of TIMSS data may also be better tailored to different national samples, and should perhaps not be applied universally across countries and educational contexts.

Comparing the DINA model to other, less parsimonious models within the DINA family, Yamaguchi and Okada (2018) reported that, in terms of model fit to TIMSS mathematics data, both the DINA model and the ‘deterministic inputs noisy OR gate' (DINO) model were outclassed by alternative sub-models in the wider generalised DINA family that make neither explicitly conjunctive or disjunctive assumptions about the interactions of different attribute masteries on item performance. Weaker model fit of the DINA models relative to models with more relaxed assumptions about attribute structures is a somewhat typical finding of these kinds of comparative CDM analyses (de la Torre, 2012). However, these more generalised models typically suffer from poorer interpretability and can be much more computationally demanding. This arguably limits their practical implementation in ‘real’ cognitive diagnostic assessments (Ravand & Robitzsch, 2015).

6.3 Methods

The analytic approach in this paper follows a similar methodology to previous studies that have employed the DINA model or similar CDMs to evaluate international assessment data (e.g. Lee et al., 2011; Hou et al., 2013; Zhao, 2013; George & Robitzsch, 2018; Yamaguchi & Okada, 2018;

Wang & Qiu, 2019, Ma & de la Torre, 2020a). However, rather than focusing on what diagnostic information is provided by these models, this paper focuses on how and why applying these models to the datasets from international assessments are likely to yield poor results and/or violate key assumptions or conditions of those models. Each issue is raised in turn, and where appropriate, an empirical demonstration is presented to highlight how this issue can be seen with regards to data from TIMSS mathematics. This section briefly discusses the test-items and samples used for these demonstrations, as well as the attributes included in two different Q-matrices used under the DINA model.

6.3.1 Data and test-items

The analyses in this paper utilise data from the fourth-grade mathematics assessment conducted as a part of TIMSS 2011. These data are publicly available from the TIMSS & PIRLS website (https://timssandpirls.bc.edu/). TIMSS was selected over other international assessments for three main, albeit likely related, reasons. Firstly, TIMSS has historically released more information about the content of their test-items compared to PISA, and this has been more explicitly linked to their published assessment frameworks. Secondly, TIMSS items are typically more independent of one another relative to the test-items used in PIRLS reading assessments, which are nested within specific texts. Thirdly, TIMSS has been used for the majority of CDM-research involving international assessment datasets (see paper 1, Chapter 4), possibly because of the first two reasons.

While the datasets of three more cycles of TIMSS data have been released since this study (TIMSS 2015, 2019 and 2023 respectively), the IEA (the International Association for the Evaluation of Educational Achievement, who are responsible for the design and development of TIMSS and PIRLS studies) have released considerably less information about the content of most of the test-items used in more recent studies. This is because, in each cycle of the study, the IEA retain most of their test-items for use in future iterations of the study. This allows for the estimation of item parameters for newly administered items to be calibrated with the estimates of

retained items that have been used in previous cycles, allowing for longitudinal comparisons of performance on a time invariant scale.

In TIMSS 2011, each student was randomly assigned one of fourteen booklets which contained two consecutively numbered item blocks (e.g. M01 and M02, M06 and M07, M14 and M01), which each block consisting of around 10-15 items. This therefore means that not all students were administered the same items. However, the distribution of items across blocks, and blocks across booklets ensured that the assessment as a whole covered the entirety of the breadth of the TIMSS mathematics assessment framework (e.g. content areas and more specific mathematics topics) in sufficient depth while minimising the demands placed on individual students. The overlapping block structure across booklets also means that two students who answered entirely different questions can still be compared. This approach is appropriate for TIMSS, as it is primarily concerned with estimating performance at national levels, rather than at the student-level.

A total of 85 test-items distributed across item blocks M01 – M07 are included in the analyses presented here. This represents around half of the items included in the study. The content of the items in blocks M01 – M03, and M05 – M07 were released after the publication of results in TIMSS 2011. Items from block M04 were discontinued after the release of results from TIMSS 2015. Access to the content of the items in block M04 was made possible upon request after the release of results from TIMSS 2015. The 85 items across blocks M01 – M07 represent the largest possible contiguous chain of released item blocks to date. However, individual students included in the sample have, at most, responses to items contained within two of these seven item blocks (~25 items), while some only have valid responses to items within one block.

TIMSS 2011 included a roughly 50:50 split of multiple-choice and constructed-response items. Of those included in the analyses presented here, 42 were multiple-choice, and 43 were constructed-response. Multiple-choice items had one correct response and three incorrect distractors, and were all scored out of one. Seven of the constructed-response items included in this analysis were originally scored out of two, with partially correct responses being awarded one

mark. As the DINA model requires all items to be scored dichotomously, polytomously scored items were recoded such that only fully correct responses were awarded a mark.

6.3.2 Student sample

Data from Australia's national sample from TIMSS 2011 was selected for these analyses. Australia was selected as an example country for three main reasons. Firstly, students were assessed in English, facilitating the discussion of specific items. Secondly, Australia's overall average score in TIMSS 2011 (516) was close to (albeit, significantly higher than) the international average score (500). Thirdly, Australia's TIMSS 2011 cohort was larger than other countries with similar characteristics, such as Ireland or New Zealand.

Data from 3,497 students were included in these analyses. This represented slightly over half of Australia's total cohort for TIMSS 2011. Those omitted from this analysis include students who were assigned booklets numbered 8 – 13, and therefore did not have any valid performance data for items within the first seven numbered item blocks (M01 – M07). Other students without valid performance data for items in these blocks, such as those who only had valid response data to items from other blocks were also omitted. This final analysis cohort did not significantly differ from Australia's overall cohort in terms of overall TIMSS mathematics scores (517 vs. 516), or in terms of performance on any of the individual mathematics content subscales as reported in TIMSS 2011.

6.3.3 Analysis and Q-matrices

Data was cleaned and analysed in R. DINA analyses were conducted using the GDINA package (Ma & de la Torre, 2020b) and focused on item-level analyses under two different Q-matrices. Q-matrix 1 follows the structure of the content domains and cognitive domains used to score performance in TIMSS. Each item included in TIMSS 2011 belonged to one of three content domains (number, geometry, and data) and one of three cognitive domains (knowing, applying, and reasoning), as defined by the TIMSS 2011 assessment framework (Mullis et al., 2009). Given that the test-items used in TIMSS 2011 were designed with these content and cognitive domains in mind, it is a sensible first step to look at estimating domain mastery under this framework.

While it would be possible to test mastery of a total of six attributes (three content domains, three cognitive domains), this would mean that every item measures exactly two attributes. To assist in creating completeness of the Q-matrix (Gu & Xu, 2021), Madison and Bradshaw (2015) suggest creating composite attributes that represent the interaction of the attributes within each subgroup. In this case, the composites are of the interaction between the content and the cognitive domain measured by each item; this mirrors the approach used by George and Robitzsch (2018) in their application of the DINA model to TIMSS data. The first Q-matrix used in this analysis, Q-matrix 1, therefore consists of nine attributes. The distribution of items by the content and cognitive domains is listed in Table 24. The full Q-matrix for all 85 items is presented in Appendix C.

Table 24: Counts of interactions of content and cognitive domains in Q-matrix 1

	Cognitive Domains			
Content domains	*Knowing*	*Applying*	*Reasoning*	∑
Number	15	19	12	46
Geometry	12	13	3	28
Data	4	4	3	11
∑	31	36	18	85

Even prior to conducting analysis, there is reason to suspect that this Q-matrix may perform poorly. In a previous study looking to predict the difficulties of this set of 85 test-items based on item features and properties, Stiff (see paper 2, Chapter 5) showed that the basic content and cognitive domains alone explained just around a quarter of the variance in item difficulties. This means that the majority of variance in item difficulties were unaccounted for by the content and cognitive domains. In other words, these results suggest that Q-matrix 1 may be heavily misspecified for cognitive diagnostic modelling purposes because it does not adequately account for what the test-items are actually testing. He further showed that an alternative framework based on Tatsuoka et al.'s (2004) early cognitive diagnostic modelling of eighth grade TIMSS mathematics fared considerably better, explaining just over half of the variance in item difficulties. The item-level features included in this analysis were split into three groups; content attributes, process attributes, and skill attributes.

As a point of comparison, this paper will also therefore explore the DINA model under a modified version of the best-fitting Q-matrix used in Stiff's (see paper 2, Chapter 5) analysis, which consisted of 19 level item-features. The modified version presented in this paper consists of 12 attributes that were found to have statistically significant unique contributions in explaining variance in item difficulties in Stiff's (see paper 2, Chapter 5) analysis and that represent suitable attributes for cognitive diagnostic modelling. A summary of these attributes, and the number of test-items measuring each attribute is presented in Table 25. Each item was specified to have assessed one of the four content attributes, and any number of the process or skill attributes ($M = 2.6, range\, 1-6$ attributes). The full assignment of attributes for Q-matrix 2 is presented in Appendix C.

Table 25: Summary of attributes in Q-matrix 2

Attribute	Description	No. of items
	Content attributes	
j_{201}	Mathematical concepts and operations involving whole numbers, fractions or decimals	42
j_{202}	Geometry content	26
j_{203}	Data-related content	10
j_{204}	Measuring or estimating in real-world contexts: (e.g. lengths, time, temperature etc.)	7
	Process attributes	
j_{205}	Perform computations in arithmetic / geometry	28
j_{206}	Make judgements or evaluate mathematical correctness	20
j_{207}	Reason logically in verbal and/or visual contexts	21
	Skill attributes	
j_{208}	Question requires a demonstration of number sense / place-value awareness	12
j_{209}	Question requires the student to identify and understand numerical patterns and sequences.	8
j_{210}	Question requires the student to identify and understand proportional relationships.	7
j_{211}	Question requires the student to compare the properties of different entities in the item.	23
j_{212}	Question requires the student to check/evaluate from a set of options (not just multiple-choice, requires active selection of most appropriate option etc.)	21

6.4 Issues

6.4.1 Q-matrix identifiability

The structure of the TIMSS dataset presents some initial complications. The analyses presented here were conducted on a set of 85 items. By design of the study however, individual students only gave responses to around 20 – 25 of these items, and in some cases, less.

One of Gu and Xu's (2021) three recommended conditions for Q-matrix identifiability, repetition, suggests that each attribute be measured by at least three test-items. Both of the overall Q-matrices meet this condition, albeit closely. However, given the TIMSS booklet structure, this is not the case for the number of responses at the student level. As an example, Table 26 shows Q-matrix 1 for students who were given booklet 1, which consisted of items from blocks M01 and M02. These students responded to a total of 21 test-items. Of the nine attributes of Q-matrix 1, only three attributes (number*applying, number*reasoning, and geometry*knowing) were adequately measured for these students (> 3 items measuring those attributes). Perhaps even more alarmingly, two attributes (geometry*reasoning and data*applying) were not measured at all by this subset of items, meaning that there is no way for the model to estimate the mastery of these attributes for these students. Similar issues are observed for different attributes across the seven booklets included in the analysis. This particularly applies to the data-related attributes, as this content domain was generally less well represented among the items included in the fourth-grade mathematics assessment of TIMSS 2011.

Table 26: Q-matrix 1 (booklet 1 only)

Item	j_{101}	j_{102}	j_{103}	j_{104}	j_{105}	j_{106}	j_{107}	j_{108}	j_{109}
i_1	0	1	0	0	0	0	0	0	0
i_2	0	0	1	0	0	0	0	0	0
i_3	0	0	1	0	0	0	0	0	0
i_4	0	0	1	0	0	0	0	0	0
i_5	0	0	1	0	0	0	0	0	0
i_6	0	1	0	0	0	0	0	0	0
i_7	0	0	0	1	0	0	0	0	0
i_8	0	0	0	1	0	0	0	0	0
i_9	0	0	1	0	0	0	0	0	0
i_{10}	0	1	0	0	0	0	0	0	0
i_{11}	1	0	0	0	0	0	0	0	0
i_{12}	0	0	1	0	0	0	0	0	0
i_{13}	0	0	1	0	0	0	0	0	0
i_{14}	1	0	0	0	0	0	0	0	0
i_{15}	0	1	0	0	0	0	0	0	0
i_{16}	0	0	0	1	0	0	0	0	0
i_{17}	0	0	0	0	1	0	0	0	0
i_{18}	0	0	0	0	1	0	0	0	0
i_{19}	0	0	0	1	0	0	0	0	0
i_{20}	0	0	0	0	0	0	1	0	0
i_{21}	0	0	0	0	0	0	0	0	1
∑	**2**	**4**	**7**	**4**	**2**	**0**	**1**	**0**	**1**

j101 = number*knowing, j102 = number*applying, j103 = *number*reasoning, j104 = geometry*knowing, j105 = geometry*applying, j106 = geometry*reasoning, j107 = data*knowing, j108 = data*applying, j109 = data*reasoning

Q-matrix 2 deals with attribute repetition slightly better than Q-matrix 1. The Q-matrix for the same subset of items included in booklet 1 is shown in Table 27. While each item in Q-matrix 1 only measured a single attribute at a time, Q-matrix 2 has high levels of within-item multidimensionality of attributes; that is, most of the test-items assess more than one attribute. This has both positive and negative implications. Generally, most students had item-level response data for all twelve attributes, though attribute j_{10}, proportional relationships, was not evenly distributed across the booklets, and so was not measured for all students. However, while Q-matrix 1 only assigned a single attribute to each test-item, 78 of the test-items in Q-matrix 2

featured more than one attribute. While this is beneficial in creating better attribute coverage, it presents an alternative problem – Q-matrix 2 does not meet another recommendation condition of Q-matrix identifiability - completeness (Gu & Xu, 2021). There are only seven items that measure a single attribute, and each of these measure one of the four content attributes (j_{101} - j_{104}). There are no test-items that measure any of the process or skill attributes of Q-matrix 2 in isolation. Gu and Xu (2021) argued that CDMs like the DINA model require each attribute to be measured in isolation by at least one item, as this helps to disentangle the effects of masteries of other attributes from one another. These issues are again compounded by the structure of the TIMSS dataset, with some students not having responses to a single item that measures a single attribute under Q-matrix 2. While it is certainly better to at least have some response data to all attributes than to have no attribute responses at all, parameter estimation is likely to be skewed if the model struggles to link item-level success or failure to individual attributes of the Q-matrix.

Table 27: Q-matrix 2 (booklet 1 only)

Item	j_{201}	j_{202}	j_{203}	j_{204}	j_{205}	j_{206}	j_{207}	j_{208}	j_{209}	j_{210}	j_{211}	j_{212}
i_1	1	0	0	0	0	0	0	0	0	1	0	0
i_2	1	0	0	0	0	0	0	0	0	1	0	0
i_3	1	0	0	0	0	1	0	0	0	1	0	0
i_4	1	0	0	0	0	0	0	0	0	1	0	0
i_5	1	0	0	0	0	0	1	0	0	1	0	0
i_6	1	0	0	0	1	0	0	0	0	0	0	0
i_7	0	1	0	0	0	0	1	0	0	0	1	1
i_8	0	1	0	0	1	0	0	0	0	0	1	0
i_9	0	0	0	1	0	0	0	0	0	1	0	0
i_{10}	0	0	0	1	1	0	0	0	0	0	0	0
i_{11}	1	0	0	0	0	1	0	1	0	0	1	0
i_{12}	1	0	0	0	0	0	1	0	0	0	0	0
i_{13}	0	0	0	1	0	0	1	0	0	0	0	0
i_{14}	1	0	0	0	1	0	0	0	0	0	0	0
i_{15}	1	0	0	0	0	0	1	0	1	0	0	0
i_{16}	0	1	0	0	0	0	0	0	0	0	1	0
i_{17}	0	1	0	0	0	0	0	0	0	0	1	0
i_{18}	0	1	0	0	1	1	0	0	0	0	0	0
i_{19}	0	1	0	0	0	1	0	0	0	0	0	0
i_{20}	0	0	1	0	0	1	0	0	0	1	0	0
i_{21}	0	0	1	0	1	0	1	0	0	0	1	0
∑	**10**	**6**	**2**	**3**	**6**	**5**	**6**	**1**	**1**	**7**	**6**	**1**

j201 = number, j202 = geometry, j203 = data, j204 = measurement and estimation, j205 = computations, j206 = judgements, j207 = reasoning, j208 = number sense, j209 = patterns and sequences, j210 = proportional relationships, j211 = comparing entities, j212 = checking options

The structure of the TIMSS dataset does not necessarily make it impossible to construct a Q-matrix that meets the conditions of identifiability, both at the overall test-level and at the booklet level. It could of course be argued that neither Q-matrix 1 nor Q-matrix 2 selected for these analyses include the correct or most suitable attributes. An alternative Q-matrix could potentially have avoided these issues. However, constructing a suitable Q-matrix that works at the booklet level and accurately reflects what the test-items measure while also meeting the identifiability conditions is challenging because the distribution of test-item content is not even across booklets.

An alternative approach would be to construct a Q-matrix that is only applied to a set of items in a single booklet. This is an approach that has been adopted by many previous articles that have applied the DINA model to TIMSS data (e.g. Lee et al., 2011; Zhao, 2013; Yamaguchi & Okada, 2018). This approach comes with other caveats. Including more items in the DINA allows for better control of item-level noise when there is less confidence in how attributes reflect test-item content. This particularly applies when the attributes used are very general. The attributes in Q-matrix 1 in arguably very broad in nature – this means that two items that share the same attribute can, in fact, look very different and are likely to be experienced very differently to examinees (Deonovic et al., 2019). Conclusions drawn about the nature of those attributes can therefore vary wildly depending on which specific items are selected for analysis. Including the items from across seven blocks of items, rather than across just one allows for some mitigation of this noise.

Sample size reduction must also be considered; one of the perceived benefits of using data from international assessments such as TIMSS or PISA for cognitive diagnostic modelling is that the sample sizes are relatively large. However, given the booklet structure, response rates at the item-level within each participating country are not particularly high. The dataset used for analysis here consists of 3,497 students from Australia, but each test-item only has responses from around 1,000 students. While it is argued that there is no strict rule on a minimum number of examinees (Rupp, 2023), simulation studies by Şen and Cohen (2021) suggested that at least 500 examinees are required as for cognitive diagnostic modelling, with sample sizes of around 1,000 needed to “adequately recover all model parameters” (p.14). They argued however that this number is likely to rise when more general CDMs are used, and when the number of attributes increases beyond three. While the DINA model makes very specific assumptions about attribute mastery, and therefore can tolerate lower-end sample sizes, the Q-matrices in this analysis include many more attributes. Including data from more test-items in TIMSS means the inclusion of more examinees, and therefore helps to stabilise parameter estimation.

6.4.2 Attribute profile misclassification

Abnormal results can easily be identified by looking at most common profiles that examinees were classified into. Under Q-matrix 1, more than a third of students were classified as having mastered none of the nine attributes. Conversely, just under a third of students were classified as having mastered all nine attributes. While it is not uncommon for the DINA model to produce somewhat high estimates for the complete mastery and complete non-mastery latent profiles (de la Torre, 2012), the pattern observed here is relatively extreme. The twenty most common attribute profiles under Q-matrix 1 are presented in Table 28.

Table 28: Most common attribute profiles (Q-matrix 1)

Attribute profile	%	Attribute profile	%
000000000	34.4%	000110011	1.0%
111111111	30.0%	000000101	1.0%
111110111	2.6%	110010111	1.0%
000000001	1.8%	010100110	0.8%
100110111	1.3%	100010100	0.8%
000000111	1.2%	110111100	0.7%
111011011	1.2%	000001111	0.6%
000000110	1.2%	000100100	0.6%
000110011	1.1%	100110001	0.6%
000000110	1.0%	000001101	0.6%

Attributes in order: number*knowing, number*applying, number*reasoning, geometry*knowing, geometry*applying, geometry*reasoning, data*knowing, data*applying, data*reasoning

Such a pattern would make sense if this were reflected by the proportion of examinees giving all-correct, or all-incorrect responses. However, only around 1% of the students in Australia's sample answered all of their test-items correctly, and even fewer failed to provide a single correct response. Given the assumptions of the DINA model, we should expect that this would be reflected in greater proportions of students with mixed attribute profiles. As this is not the case, it strongly suggests that the previously discussed issues with the Q-matrix are producing skewed mastery estimation.

A slightly different situation emerges under Q-matrix 2. While the complete mastery profile once again is overrepresented, there is a substantial drop-off in the proportion of students who were estimated to have not mastered any of the attributes. This is likely due to most items in Q-matrix 2 being classified as targeting multiple attributes. The twenty most common attribute profiles under Q-matrix 2 are reported in Table 29.

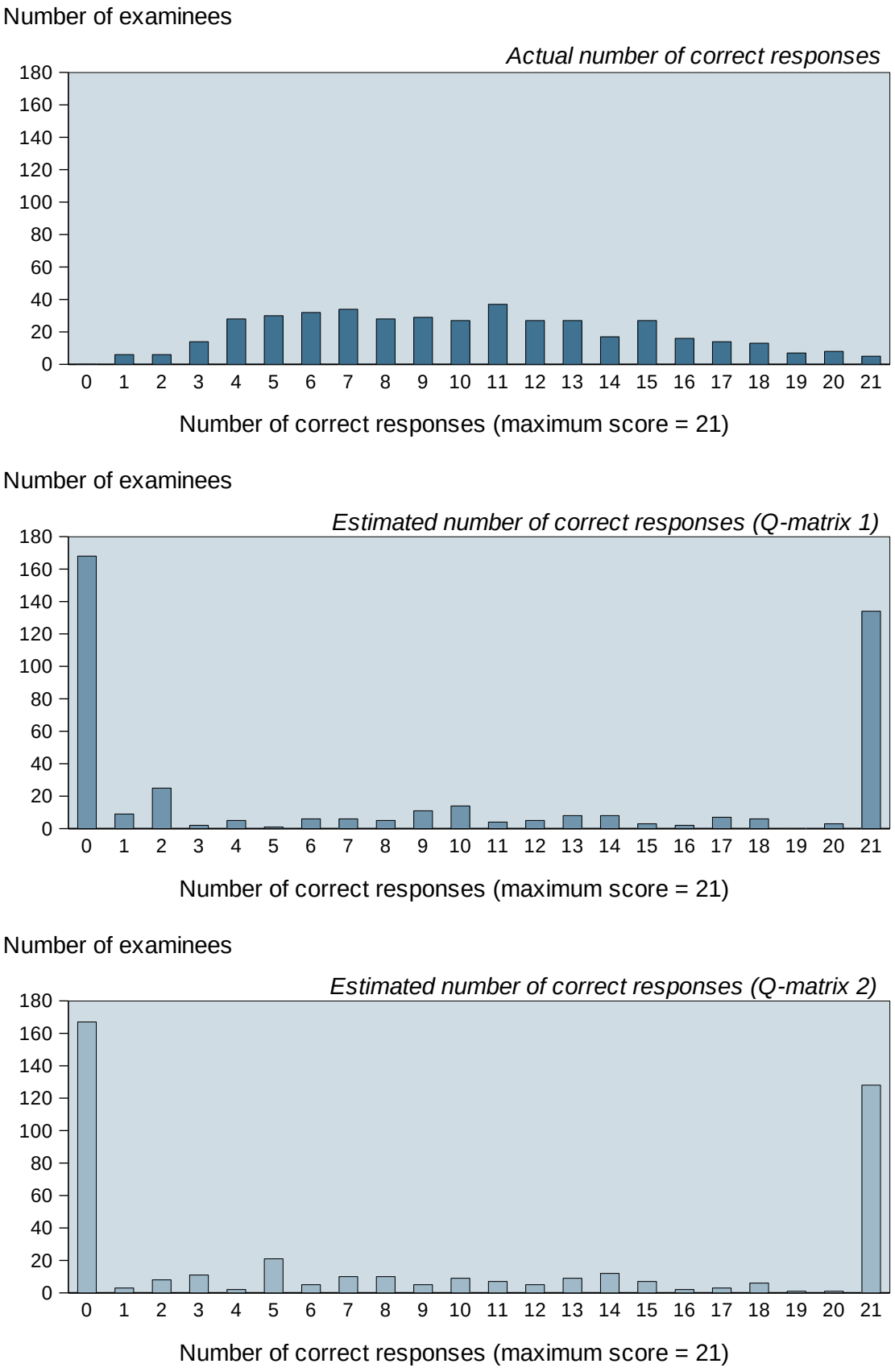
Table 29: Most common attribute profiles (Q-matrix 2)

Attribute profile	%	Attribute profile	%
111111111111	31.7%	000000001100	0.6%
000000000000	2.0%	000100000000	0.6%
111111111011	1.8%	001000000100	0.6%
000000000100	1.5%	000000010000	0.6%
011101101111	1.3%	101111011111	0.6%
111001111111	0.9%	111111110111	0.6%
110001001101	0.9%	111011111111	0.5%
000000001000	0.8%	111011110111	0.5%
001000000000	0.8%	111111010111	0.5%
000000000001	0.6%	110010110111	0.5%

Attributes in order: number, geometry, data, real-world measurement and estimation, computations, judgements, reasoning, number-sense, patterns and sequences, proportional reasoning, comparing entities, evaluating options

This issue is represented graphically in Figure 16, which focuses on the item-level performance of students who were given booklet 1. The actual distribution of scores (out of a possible 21) is somewhat normally distributed, and reflects the distribution of TIMSS scores achieved by Australian students in TIMSS 2011. This distribution (represented in the top panel) contrasts starkly to that of the expected distributions of item-level performance based on the mastery profiles estimated by Q-matrices 1 and 2, in which the overwhelming majority of students are estimated to have answered all or no questions correctly. Similar distributions are observed across all seven booklets included in this analysis.

Figure 16: Distributions of actual and expected item-level performance (booklet 1)



6.4.3 Item-level guessing and slipping parameters

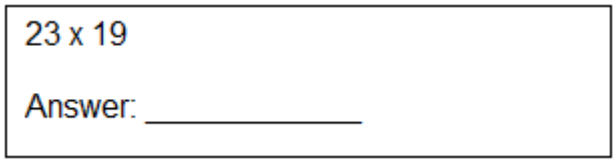
Issues with poor model fit can also be observed when inspecting individual item-level parameters. One initial check is whether any items violate the assumption of monotonicity; that is, that an increase in an examinees' ability level (in this case, as demonstrated by attribute mastery) has a negative association with item-level success. A check for this can be made by comparing item-level guessing and slipping parameters. Monotonicity violations can be observed if there are items where $1-s_i<g_i$ (Xu et al., 2016). This assumption was not violated for any items under either Q-matrix 1 or Q-matrix 2.

Many test-items included in this analysis exhibited either very extreme guessing or slipping parameters. A full summary of these is provided in Appendix F. However, the general pattern was for more difficult test-items to have larger slipping parameters, and easier items to have larger guessing parameters, with the DINA model performing best on test-items with middling difficulty levels. In other words, the DINA model's results suggest that large proportions of examinees underperformed relative to expectation on high difficulty items (thus, 'slipping'), while large proportions of examinees were able to produce the correct answers to easy items, despite lacking the requisite skill masteries (thus, 'guessing'). This is an unsurprising finding given that the easiest and most difficult items are likely to deviate most from the expected pattern, especially if the Q-matrix does not sufficiently account for why those items are more or less difficult than others. Given how extreme many of these slipping and guessing parameters were at the item-level, these results most likely indicate that the model was poor at estimating examinees' actual mastery and non-mastery of the attributes in Q-matrix 1 and Q-matrix 2.

Figure 17 presents an example of a test-item which was found to have very large slipping parameters. Under Q-matrix 1, it was categorised as testing the 'number*knowing' attribute, while under Q-matrix 2, it was categorised as testing two attributes – 'number', and 'performing computations'. Under both Q-matrices, almost 80% of examinees who were estimated to have mastered the required attribute(s) failed to provide the correct response ($s_i=.79$ and $s_i=.78$ respectively). By contrast, only 4% and 3% of students who were not estimated to possess the

required masteries provided the correct response (g_i=.04 and g_i=.03 respectively). Only 12% of the students in Australia's sample were actually able to work out the correct answer to this item.

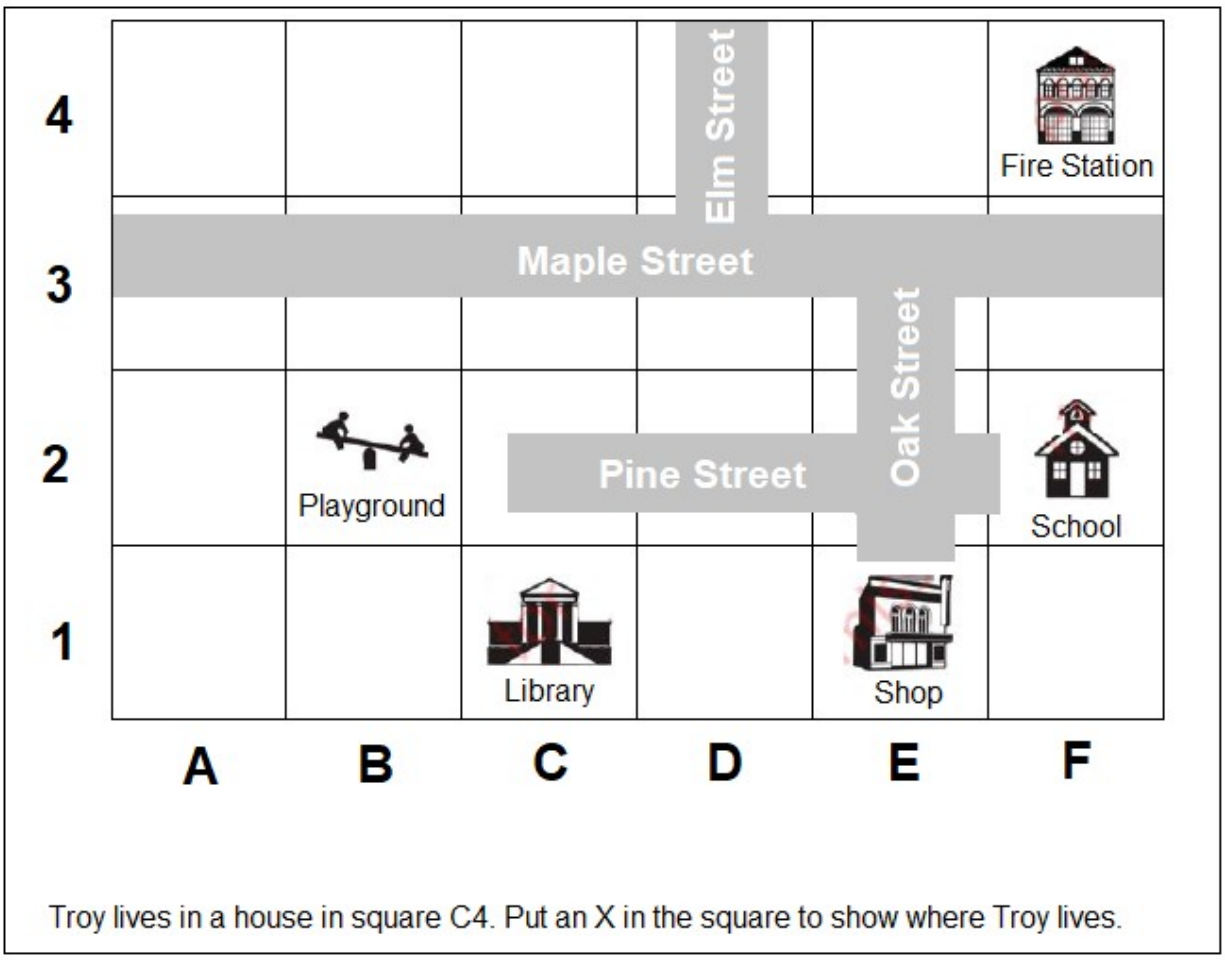
Figure 17: Example item 14 with high slipping parameter estimates



Source: TIMSS 2011 Assessment. Copyright © 2013 International Association for the Evaluation of Educational Achievement (IEA). Publisher: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

In contrast, Figure 18 presents an example of an item with very high guessing parameters. Under Q-matrix 1, it was categorised as testing the 'geometry*applying' attribute, and under Q-matrix 2, it was categorised as testing two attributes – 'geometry' and 'comparing entities'. Under both Q-matrices, more than 90% of students who were profiled to lack the requisite attributes correctly placed the X in square C4 (g_i=.94 under both Q-matrices). Only 3% of the students in Australia's TIMSS 2011 sample gave an incorrect response to this test-item, indicating that the Q-matrices fail to account for the easiness of the item.

Figure 18: Example item 14 with high guessing parameter estimates



Source: TIMSS 2011 Assessment. Copyright © 2013 International Association for the Evaluation of Educational Achievement (IEA). Publisher: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

One way of evaluating the accuracy of the DINA model's attribute estimation is by calculating item-level discrimination. This can be computed as $1-g_i-s_i$. This represents the overall proportion of examinees whose responses conform to the expectations of the DINA model (i.e. those who are estimated to have neither incorrectly slipped or correctly guessed). Guessing and slipping are both normal and expected examinee behaviours, but high values of either indicate poor model fit; high values of *both* suggest problems either with the test-item itself, and/or Q-matrix misspecification.

The test-items under Q-matrix 1 had similar mean-item level discriminations, with similar ranges (Q-matrix 1; $M=.41, range=.05-.74$; Q-matrix 2 $M=.42, range=.05-.72$). This suggests that the two Q-matrices were equally poor. This is despite Q-matrix 2 being developed around a Q-matrix that in a previous study, relative to Q-matrix 1, accounted for a significantly greater proportion of the variance in item difficulties (see paper 2, Chapter 5). The guessing and slipping

parameters for each item might therefore have been predicted to be lower in Q-matrix 2 relative to Q-matrix 1. This could be because the modifications made to that Q-matrix to suit cognitive diagnostic modelling may have weakened the Q-matrix's relationship with item difficulties. While this is possible (and indeed, around 50% of the variance in item difficulties was still unaccounted for in that previous analysis), it is also likely that the lack of Q-matrix identifiability (as discussed in section 6.4.1) heavily contribute to the problems in validly estimating students' mastery profiles for both Q-matrices.

6.5 Discussion

This paper looked to identify and demonstrate potential issues in the application of CDMs to item-level performance data from international assessments. Data from the fourth-grade mathematics assessment of TIMSS 2011 was selected to help demonstrate these issues in practice.

Results at both the examinee-level and item-level exhibit poor fit to the underlying data. This is evident from the extreme patterns of attribute mastery and non-mastery, and high item-level guessing and slipping parameters. There are many potential causes of poor model fit. In some contexts, this could be used as evidence that the items themselves are poorly designed and do not relate well to examinee ability. As the items used in assessments such as TIMSS undergo thorough pilot testing across multiple countries, and poorly functioning items are removed from these assessments altogether, it is more likely that the poor fit reported in these analyses relate to issues with the modelling instead.

In comparative analyses of various CDMs, the DINA model is typically not the preferred model when attempting to maximise fit to the data (de la Torre, 2012; Ravand & Baghaei, 2020; Williamson, 2023; Yamaguchi & Okada, 2018). Generalised models typically outperform saturated models like the DINA, in part because they make fewer assumptions about how mastery of different attributes interact with one another to shape and predict item-level success (Deonovic et al., 2019). However, these more generalised models are more complex to interpret, require larger sample sizes, and more likely to suffer issues with Q-matrix identifiability (Williamson, 2023). For the purpose of these demonstrations, the DINA model was selected as

the model of choice, but it should be noted that some of these issues could potentially have been mitigated with alternative model selection.

However, adopting a different CDM would not have tackled the core problem with the application of CDMs to international assessment data. PISA and PIRLS are structurally very similar to TIMSS, and share similar goals; these studies are not necessarily designed to be diagnostic at the student-level. Rather, their focus is on drawing broader, national conclusions about educational standards within and across different countries. This is reflected in the way in which the tests are administered. They are also concerned with assessing subject performance broadly (e.g. in mathematics, science and reading) rather than at the level of granularity that CDMs were designed to evaluate. It is arguably not surprising that previous research has utilised the easily accessible, large datasets offered by these studies for cognitive diagnostic modelling purposes, but the scope of the assessed content in these studies is simply too broad.

In all cognitive diagnostic modelling studies, there is a necessary trade-off between attribute granularity and test content breadth. This arises because of the need for there to be sufficient coverage of each attribute across the test to allow for valid estimation of examinee's masteries (Gu & Xu, 2021). Students who participate in international assessment studies have a finite number of responses to items covering wildly different content areas. Given the breadth of topics covered in TIMSS, for example, it is not possible to construct a Q-matrix that ensures sufficient repetition of attributes across items without relying on very high-level attributes such as 'number' and 'geometry', and as demonstrated, attribute coverage is still problematic at the examinee-level. By contrast, most simulation studies or theoretical discussions of the application of CDMs focus on much more granular topics, such as computations with fractions (e.g. Tatsuoka, 1990), where attributes can distinguish between much more specific and relevant skills (e.g. 'finding a common denominator' or 'converting a whole number to a fraction'). In addition to the validity issues relating to Q-matrix construction for international assessment datasets, the wider question must be raised as to how pedagogically useful the findings of such models even are when applied to these datasets.

There have been many challenges in the real-world applications of cognitive diagnosis assessment (Ravand & Baghaei, 2020). CDMs demand large sample sizes, careful item construction and complex Q-matrix validation. The datasets from international assessment superficially appear as though they may be well suited to these kinds of analysis and could be used to improve the reach and impact of CDMs. However, there is a fundamental mismatch between the goals and intentions of international large-scale assessments with those of cognitive diagnostic assessment that is reflected in the content of their test-items. It is the argument of this paper that the results of cognitive diagnostic modelling studies of international assessment data are likely meaningless and should be interpreted with a healthy level of scepticism.

6.6 References

Addey, C. (2017). Golden relics & historical standards: how the OECD is expanding global education governance through PISA for Development. *Critical Studies in Education, 58:3*, 311-325, https://doi.org/10.1080/17508487.2017.1352006

Chen, Y.H. (2011). Cognitive diagnosis of mathematics performance between rural and urban students in Taiwan. *Assessment in Education: Principles, Policy & Practice, 19(2)*, 193-209. https://doi.org/10.1080/0969594X.2011.560562

Choi, K.M., Lee, Y.S., & Park, Y.S. (2015). What CDM can tell about what students have learned: an analysis of TIMSS eighth grade mathematics. *Eurasia Journal of Mathematics, Science and Technology Education, 11(6)*, 1563-1577. https://doi.org/10.12973/eurasia.2015.1421a

de la Torre, J. (2012). The generalized DINA model framework. *Psychometrika, 76*, 179-199. https://doi.org/10.1007/s11336-011-9207-7

de la Torre, J., & Douglas, J.A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69(3)*, 333-353. https://doi.org/10.1007/BF02295640

Deonovic, B., Chopade, P., Yudelson, M., de la Torre, J., & von Davier, A.A. (2019). Application of cognitive diagnostic models to learning and assessment systems. In: M. von Davier & Y.S. Lee (Eds.), *Handbook of diagnostic classification models: Models and model extensions, applications, software packages* (pp. 437-460). Springer.

Fischman, G.E., Topper, A.M., Silova, I., Goebel, J., & Holloway, J.L. (2018). Examining the influence of international large-scale assessments on national education policies. *Journal of Education Policy, 34(4)*, 470-499. https://doi.org/10.1080/02680939.2018.1460493

George, A.C., & Robitzsch, A. (2018). Focusing on interactions between content and cognition: a new perspective on gender differences in mathematical sub-competencies. *Applied Measurement in Education, 31(1)*, 79-97. https://doi.org/10.1080/08957347.2017.1391260

Gierl, M.J., & Cui, Y. (2008). Defining characteristics of diagnostic classification models and the problem of retrofitting in cognitive diagnostic assessment. *Measurement: Interdisciplinary Research and Perspectives, 6(4),* 263-268. https://doi.org/10.1080/15366360802497762

Gorin, J.S. (2009). Diagnostic classification models: are they necessary? Commentary on Rupp and Templin (2008). *Measurement: Interdisciplinary Research and Perspectives, 7(1),* 30-33. https://doi.org/10.1080/15366360802715387

Gorur, R. (2017). Statistics and statecraft: exploring the potentials, politics and practices of international educational assessment. *Critical Studies in Education, 58(3)*, 261-265. https://doi.org/10.1080/17508487.2017.1353271

Gu, Y., & Xu, G. (2021). Sufficient and necessary conditions for the identifiability of the Q-matrix. *Statistica Sinica, 31(1)*, 449-472. https://doi.org/10.5705/ss.202018.0410

Hansen, M., Cai, L., Monroe, S., & Li, Z. (2014). Limited-information goodness-of-fit testing of diagnostic classification item response theory models. *British Journal of Mathematical and Statistical Psychology, 69, 3,* 225-252. https://doi.org/10.1111/bmsp.12074

Holliday, W.G., & Holliday, B.W. (2003). Why using international comparative math and science achievement data from TIMSS is not helpful. *The Educational Forum, 67*, 250–258. https://doi.org/10.1080/00131720309335038

Hou, L., Nandakumar, R., & de la Torre, J. (2013). Differential item functioning assessment in cognitive diagnostic modeling: application of the Wald test to investigate DIF in the DINA model. *Journal of Educational Measurement, 51*, 98-125. https://doi.org/10.1111/jedm.12036

Jang, E.E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: validity arguments for fusion model application to LanguEdge assessment. *Language Testing, 26(1)*, 31-73. https://doi.org/10.1177/0265532208097336

Junker, B.W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25(3)*, 258-272. https://doi.org/10.1177/01466210122032064

Lee, Y.S., Park, Y.S., & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the U.S. national sample using the TIMSS 2007. *International Journal of Testing, 11(2)*, 144-177. https://doi.org/10.1080/15305058.2010.534571

Lee, Y.W., & Sawaki, Y. (2009). Cognitive diagnosis approaches to language assessment: An overview. *Language Assessment Quarterly, 6(3)*, 172-189. https://doi.org/10.1080/15434300902985108

Leighton, J., & Gierl, M. (Eds.). (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge University Press.

Long, C. (2007). *What can we learn from TIMSS 2003*. Centre for Evaluation and Assessment, University of Pretoria. Retrieved from: https://www.amesa.org.za/AMESA2007/Volumes/Vol11.pdf

Ma, W. & de la Torre, J. (2020a). An empirical Q-matrix validation method for the sequential generalized DINA model. *British Journal of Mathematical and Statistical Psychology, 73*, 142-163. https://doi.org/10.1111/bmsp.12156

Ma, W. & de la Torre, J. (2020b). GDINA: An R package for cognitive diagnosis modeling. *Journal of Statistical Software*, *93(14)*, 1-26. https://doi:10.18637/jss.v093.i14

Madison, M.J., & Bradshaw, L.P. (2015). The effects of Q-matrix design on classification accuracy in the log-linear cognitive diagnosis model. *Educational and Psychological Measurement, 75(3)*, 491-511. https://doi.org/10.1177/0013164414539162

Mullis, I.V.S., Martin, M.O., Ruddock, G.J., O'Sullivan, C.Y. & Preuschoff, C. (2009*). TIMSS 2011 assessment frameworks*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from: https://timssandpirls.bc.edu/methods/t-instrument.html

Ravand, H., & Baghaei, P. (2020). Diagnostic classification models: recent developments, practical issues, and prospects. *International Journal of Testing, 20(1),* 24-56. https://doi.org/10.1080/15305058.2019.1588278

Rupp, A.A. (2007). The answer is in the question: A guide for describing and investigating the conceptual foundations and statistical properties of cognitive psychometric models*. International Journal of Testing, 7(2)*, 95-125. https://doi.org/10.1080/15305050701193454

Rupp, A. A. (2023). *Primer on diagnostic classification models* (2.0 ed.). Center for Assessment. https://www.nciea.org/library/primer-on-diagnostic-classificationmodels-dcms/

Rupp, A.A. (2023). Primer on diagnostic classification models (2.0 ed.). Center for Assessment. Retrieved from: https://www.nciea.org/library/primer-on-diagnostic-classificationmodels-dcms/

Tatsuoka, K.K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In: *Diagnostic Monitoring of Skill and Knowledge Acquisition* (eds N. Frederiksen, R. Glaser, A. Lesgold and M. Shafto), pp.453-488. Hillsdale: Erlbaum

Tatsuoka, K K., Corter, J.E., & Tatsuoka, C. (2004). Patterns of diagnosed mathematical content and process skills in TIMSS-R across a sample of 20 countries*. American Educational Research Journal, 41(4)*, 901-926. https://doi.org/10.3102/00028312041004901

Wagemaker, H. (2020). Study design and evolution, and the imperatives of reliability and validity. In: *Reliability and validity of international large-scale assessment: understanding IEA's comparative studies of student achievement*. Cham: Springer International Publishing.

Wang, J. (2001). TIMSS primary and middle school data: Some technical concerns. *Educational Researcher, 30,* 17-21. https://doi.org/10.3102/0013189X030006017

Wang, W.C., & Qiu, X.L. (2019). Multilevel modeling of cognitive diagnostic assessment: the multilevel DINA example. *Applied Psychological Measurement, 43(1)*, 34-50. https://doi.org/10.1177/0146621618765713

Williamson, J. (2023). *Cognitive diagnostic models and how they can be useful.* Cambridge University Press & Assessment. https://doi.org/10.17863/CAM.110853

Xu, G., Wang, C., & Shang, Z. (2016). On initial item selection in cognitive diagnostic computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology, 69*. https://doi.org/10.1111/bmsp.12072

Yamaguchi, K., & Okada, K. (2018). Comparison among cognitive diagnostic models for the TIMSS 2007 fourth grade mathematics assessment*. PloS One, 13(2),* https://doi.org/10.1371/journal.pone.0188691

Zhang, S., Liu, J., & Ying, Z. (2023). Statistical applications to cognitive diagnostic testing. *Annual Review of Statistics and Its Application, 10(1)*, 651-675. https://doi.org/10.1146/annurev-statistics-033021-111803

Zhao, F. (2013). *Using noncompensatory models in cognitive diagnostic mathematics assessments: An evaluation based on empirical data*. Doctoral dissertation, University of Kansas. Retrieved from: https://www.proquest.com/dissertations-theses/using-noncompensatory-models-cognitive-diagnostic/docview/1428738807/se-2

7 Discussion

7.1 Reflections on overarching thesis aims

This thesis investigated how cognitive psychometric models, both item-explanatory and person-diagnostic in nature, can, and have been used to analyse data from TIMSS. Specifically, it asked whether cognitive psychometric modelling approaches could be "viably and validly applied to TIMSS data to provide diagnostic information about students' knowledge and skills in mathematics". The need for such a question is warranted by the increasing amount of research applying CPMs to data from international large-scale assessments, and in particular TIMSS, despite broader concerns about the validity of CPM retrofitting analyses.

The key conclusion of this thesis is that **TIMSS is not suitable for cognitive psychometric modelling**. While the focus of the analysis has been on data from TIMSS, and to a lesser extent, PISA and PIRLS, I believe this argument is likely to apply more broadly to most retrofitting of CPMs to pre-existing assessment datasets. If there is a mismatch between the original goals of an assessment and the goals of a cognitive psychometric modelling analysis, there will almost surely be challenges and issues in the construction of a valid Q-matrix. This is because it is unlikely that the test validly measures the constructs that the retrofitting analysis wants it to have measured, and/or the volume of data (either at the test-level or the student-level) does not allow the CPM to calculate valid parameter estimates.

In the context of TIMSS, the scope of content within the mathematics assessment framework (which defines what content is included in each of the subjects and grades assessed) is too broad relative to the number of test-items. This is true at the overall test-level, and even more problematic at the student-level, given the TIMSS booklet design. This means that several compromises have to be made with the construction of the Q-matrix so as to meet the basic requirements for these models to calculate estimates for each attribute, and for person-diagnostic analyses, mastery estimates for each examinee's mastery of each attribute.

Ideally, the Q-matrix used in a CPM would consist of a small number of granular attributes representing distinct but highly related skills or knowledge states. This is typically the case in CDM simulation studies (e.g. Şen & Cohen, 2021; Paulsen & Valdivia, 2021; de la Torre & Douglas, 2004), which usually include Q-matrices of fewer than six attributes. This has also been the case with papers that have explained or introduced the methodologies of various CPMs. For example, in Fischer's (1972) presentation of the LLTM, he included an example of how the LLTM could be applied to a set of 29 differentiation problems, using seven operations as item-level factors (e.g. differentiation of a polynomial, the quotient rule, operations involving $\sin(x)$ etc.). Similarly, in Tatsuoka's (1990) introduction of the rule-space method, she presented an example of latent class modelling of performance on a set of fractions test-items, where the Q-matrix was comprised of seven highly granular attributes related to fractions.

The Q-matrices used in these demonstrations differ vastly from the applications of CPMs to TIMSS data in this thesis and in many of the articles identified in paper 1, where the item-level factors or attributes of the Q-matrices represent far broader mathematical concepts and cognitive processes. The breadth of the content coverage in TIMSS (and indeed, other large-scale international assessments) is simply too wide to create a sufficiently granular Q-matrix; the selected attributes must be far more general to ensure sufficient coverage of each attribute. Thus, there is a weaker relationship between test-item content and attributes than would be desired. I believe this is an unavoidable problem with applications of CPMs to TIMSS data.

It should be reiterated that this thesis does not directly concern itself with the evaluation of either TIMSS or any specific CPM. Rather, in this thesis, I have argued that the application of CPMs to TIMSS data, and particularly those concerned with person-diagnostic modelling, will produce nonsensical results because the Q-matrices used in these analyses will *always* be misspecified.

The following section provides some more specific reflections on findings from each of the three papers and the sub-questions each paper targeted, before returning to further reflections on the thesis' overarching question of whether TIMSS is suitable for cognitive psychometric modelling.

7.2 Reflections on findings from papers

7.2.1 Paper 1

Paper 1 addressed the following thesis sub-question:

"What are the main themes, foci, and findings of research utilising international assessment datasets for cognitive psychometric modelling purposes?"

A thematic review was conducted to address this question, targeting research articles that utilised data from any one of the three major international assessments (TIMSS, PISA, PIRLS). Of the 42 articles identified in this review, TIMSS was by far the most widely used in cognitive psychometric modelling analyses. Further, the vast majority of these had focused on person-diagnostic modelling of performance in TIMSS mathematics, rather than on item-explanatory modelling, or modelling of performance in TIMSS science.

One of the arguments of this thesis is that item-explanatory modelling should act as a precursor to person-diagnostic modelling. This is because there is a need to establish how/if the attributes selected for person-diagnostic modelling are actually measured by items. If the selected attributes for person-diagnostic modelling do not impact on the difficulty of test-items, then these attributes are unlikely to yield useful information. This is particularly pertinent for a retrofitting analysis, where the test-items have not necessarily been designed around the attributes selected for person-diagnostic modelling, and as such, there is a greater risk of Q-matrix misspecification.

The scarcity of research conducting item-explanatory modelling of TIMSS mathematics relative to person-diagnostic modelling was therefore somewhat surprising. While some articles had made direct use of TIMSS mathematics assessment frameworks (e.g. Mullis et al., 2009) in the selection of attributes for mastery diagnosis, most of the located articles took or adapted their attribute frameworks from either Tatsuoka's (2004) application of the rule-space method to TIMSS data, or from Lee et al.'s (2011) application of the DINA model, which had only taken partial inspiration from the relevant TIMSS mathematics assessment frameworks.

One of the critical reflections of this thematic review was that there was a need to empirically identify relevant attributes for cognitive diagnosis by first modelling the item-level properties that

explain the difficulties of items in TIMSS. Without reliable evidence that the attributes used in a person-diagnostic analysis align with test-item content, diagnostic CPM claims are not only statistically weak but also risk misleading researchers and policymakers about what TIMSS actually measures. This reflection helped to shape the thesis sub-questions targeted by paper 2 and paper 3.

7.2.2 Paper 2

Paper 2 addressed the following thesis sub-question:

"Does the TIMSS mathematics framework possess sufficient predictive power to explain the item-level factors contributing to item difficulty, and if not, how could the framework be improved upon?"

To address this question, the difficulties of 85 test-items from the fourth-grade mathematics assessment of TIMSS 2011 were estimated under the one-parameter logistic model (1PL). These estimates were then compared with item difficulties estimated using the linear logistic test model (LLTM), where item difficulties are estimated from the linear summation of the difficulty parameters associated with item-level attributes. Four different Q-matrices of item-level properties were used in the calculation of LLTM item difficulty estimates – three of these were taken and adapted from the TIMSS 2011 mathematics assessment framework (Mullis et al., 2009), while the fourth (Q-matrix D) was adapted from the Q-matrix used in Tatsuoka's (2004) rule-space method analysis of TIMSS mathematics data.

Unsurprisingly, the 1PL model showed significantly greater fit to the data than any of the four LLTMs. Item difficulties estimated under Q-matrix D accounted for just under half of the total variance in 1PL difficulty estimates. By comparison, Q-matrix A only explained around a quarter of the variance in 1PL difficulty estimates. Q-matrix B, which added greater detail on topic areas within the mathematics content domains of the TIMSS assessment framework only provided small improvements to the amount of explained variance, while Q-matrix C, which added information on item-level reading demands, provided no improvement at all.

This latter finding was particularly surprising given that previous research has suggested that reading demands in mathematics items are a significant contributor to item difficulty (e.g. Rewer

& Greefrath, 2023; Abedi & Lord, 2001) and has been recognised by the IEA as a potential unwanted source of item difficulty in TIMSS mathematics and TIMSS science (Martin et al., 2013). Further, of the 19 attributes included in Q-matrix D, only two were found to not provide any significant unique contributions to explained variance. The rater judged reading demand of items was one of these. There are a few possible explanations. One is that the IEA are relatively successful in minimising the reading-related demands of TIMSS mathematics items; this is a reasonable interpretation given that the word counts of most of the fourth-grade (and indeed, eighth-grade) mathematics test-items were low. Another explanation could be that more difficult test-items (be it due to the number of, or magnitude of mathematically relevant sources of item difficulty) necessitate longer word counts and/or more complex language that may increase item difficulties, and therefore reading-related contributions to item difficulties could be being masked by strong correlations with other item-level factors.

However, one argument against this interpretation is that a disproportionate number of items that were classified as having high reading demands in Q-matrices C and D were those belonging to the data content domain, which across all four Q-matrices was associated with low item difficulties. In my opinion, it is more likely that the reading-related factors used in Q-matrices C and D simply do not capture the reading-related demands of these test-items as experienced by students. Given the word-counts of these test-items, and the other issues this thesis has discussed relating to Q-matrix construction, there is limited scope for investigating reading-related contributions to item difficulty of TIMSS mathematics items, though this is an important area for investigation in mathematics assessment more broadly.

The key conclusion from this paper was that the TIMSS assessment framework, which has been the predominant framework used as the basis for Q-matrices for person-diagnostic modelling in recent studies (as reported in paper 1), does not have a strong relationship with the difficulties of TIMSS mathematics items. This is not a critique of the quality of the items used in TIMSS, nor of the conclusions drawn by the IEA about TIMSS; the IEA have not claimed that the cognitive domains, content domains, or topic areas as listed within the TIMSS mathematics assessment frameworks explain or account for the variance in item difficulties. However, this has arguably

been an implicit assumption of the cognitive diagnostic analyses that have used the TIMSS assessment frameworks as the basis of their Q-matrices.

While it may be easy to apply a Q-matrix designed around the TIMSS assessment framework to item-level performance data, the content domains and cognitive domains alone do not adequately account for the range of difficulties of these test-items. This therefore raises concerns about the validity of the findings drawn from the previous analyses identified in paper 1 that seemingly assumed the strength of this relationship.

The argument arising from this paper therefore was that future person-diagnostic modelling analyses would benefit from identifying a stronger fitting model of the difficulties of TIMSS test-items first, in order to demonstrate that the test-items used for those diagnoses actually measure the skills and knowledge states (as defined by the best fitting attributes) that are used to explain examinee performance. As paper 3 suggested however, there may be greater issues with data from TIMSS that even a Q-matrix of attributes with stronger links to item difficulty cannot overcome.

7.2.3 Paper 3

Paper 3 addressed the following thesis sub-question:

"Are the item-level data available from TIMSS suitable for cognitive diagnostic modelling of student performance, and if not, what are the threats to the validity of these analyses?"

A number of concerns associated with the retrofitting of cognitive diagnosis models such as the deterministic inputs, noisy 'and' gate (DINA) model have been raised somewhat broadly, but to my knowledge, had never been directly presented in the context of international assessment studies. This research question was therefore concerned with how these arguments relate to international assessments, and of particular relevance for this thesis, TIMSS.

To address this research question, theoretical and methodological issues associated with retrofitting to TIMSS data were discussed with reference to the results of an application of the DINA model to Australia's TIMSS 2011 fourth-grade mathematics data. The results of the

presented analyses were strongly indicative of Q-matrix misspecification. As discussed in the paper, I believe that this is a fundamental issue underlying the structure of the TIMSS assessment which imposes too many limitations on Q-matrix construction. This was discussed with reference to Gu and Xu's (2021) recommended properties for an 'identifiable' Q-matrix, which affects the extent to which accurate estimations can be made for each examinee's mastery of each attribute.

Two Q-matrices were used for the analyses presented in the paper. Both of these Q-matrices differed slightly from those used in paper 2. To avoid confusion for the purpose of this thesis, these were numbered as matrices 1-2, rather than labelled with letters. Q-matrix 1 was adapted from Q-matrix A. Rather than representing each of the content and cognitive domain attributes individually, Q-matrix 1 featured nine attributes with each representing an interaction between the content and cognitive domain for that test-item (e.g. number*reasoning). Q-matrix 2 meanwhile was adapted from Q-matrix D. The twelve selected attributes for Q-matrix 2 were either taken directly from Q-matrix D, or represented a combination of two highly-related but less well represented attributes from Q-matrix D (e.g. verbal and visual reasoning).

The sequential rationale for papers 2 and 3 was to first identify a Q-matrix of item properties and cognitive attributes that had a strong, or at least moderate predictive relationship with item difficulties, and then assess whether that better fitting Q-matrix would also work as a better fitting Q-matrix for person-diagnostic modelling. While Q-matrix D was identified as having stronger explanatory power of item difficulties than the Q-matrices taken from the TIMSS 2011 assessment framework (A-C), Q-matrix 2 performed similarly poorly to Q-matrix 1 in paper 3. Neither Q-matrix 1 nor Q-matrix 2 held all three of Gu and Xu's (2021) recommended properties for an identifiable Q-matrix at the overall test-level, nor at the booklet-level (i.e. what individual examinees responded to). Each test-item under Q-matrix 1 only assessed one of the nine attributes, meaning that there were no test-items that measured multiple attributes simultaneously. More egregiously, given the booklet structure used in TIMSS 2011 and the distribution of item content across booklets, there was insufficient repetition of attribute-level performance data for individual examinees, with most examinees having no attribute-level data

for one or more attributes. Q-matrix 2 meanwhile lacked ‘completeness’ because there were *no* items that measured a single attribute in isolation.

While I still believe that identifying the sources of difficulty of test-items is an important precursor to cognitive diagnostic modelling, one of the main conclusions drawn from paper 3 is that it is not possible to construct an ‘identifiable’ Q-matrix (Gu & Xu, 2021) that accurately represents the content of TIMSS mathematics items and is simultaneously suitable for a cognitive diagnostic modelling analysis. It is my view that no amount of prior explanatory item difficulty modelling can overcome this limitation.

7.3 Limitations

While the limitations of the analyses presented in each paper have been discussed in their relevant chapters, it is worth briefly touching on some of the broader limitations of the research presented in this thesis, particularly with respect to the arguments presented by the results of paper 2 and paper 3.

7.3.1 Model selection

One of the greatest challenges in conducting any form of cognitive psychometric modelling is model selection. As has been suggested or reported for both item-explanatory modelling and person-diagnostic modelling, the most parsimonious models are typically not the best fitting to item-level assessment data. For example, the linear logistic test model (LLTM) makes very rigid assumptions about the additive interaction of difficulty factors, rather than accounting for potentially more complex interactions between factors. Person-diagnostic models such as the DINA model and other saturated models meanwhile have fallen out of favour somewhat because they typically yield poor fit to assessment datasets relative to more general models (Ravand & Baghaei, 2020; Williamson, 2023). In papers 2 and 3, I have opted for the use of these more parsimonious models, which may be seen as a weakness in my research. There are both theoretical and practical reasons as to why I adopted these approaches.

In many of the more recent cognitive diagnostic modelling studies (both retrofitting and simulation sides), multiple CDMs are explored, and the fit of each model for each item are compared. The best fitting models can then be selected at the item-level, rather than at the test level. This approach is consistent with the belief that there are more complex relationships between attribute masteries and examinees' probabilities of producing correct answers than any one diagnostic model can account for. A proponent of this approach may therefore argue that my approach in paper 3, for example, underestimates the extent to which item and person parameters can be validly estimated from TIMSS data because of limitations in model selection. I do accept this argument. I recognise that model selection in both paper 2 and paper 3 is limited. I did not select either the LLTM or DINA models because I necessarily saw them as superior to other explanatory models, nor because I believed that their assumptions best explained item-level performance. Rather, they were chosen primarily because of their relative parsimony, and also in part due to computational limitations associated with more complex models. I fully acknowledge that, with alternative model selection, I could have explained a greater proportion of the variance in item difficulties in paper 2. Albeit with some reservations, I do also acknowledge that the poor fit presented in paper 3 could have been less extreme had I used a wider variety of CDMs than just the DINA model.

However, my main counterargument to these critiques is that it assumes poor model fit is inherently an issue of inadequate model selection, rather than an issue with either a) the Q-matrix, or b) the data itself. One of the main discussion points raised in paper 3 was that the TIMSS assessment itself is likely unsuitable for cognitive diagnostic modelling because it is not possible to construct a valid identifiable Q-matrix around TIMSS mathematics test-items. Small changes to the method of parameter estimation, either at the test-level or item-level, will not magically solve this issue. Further, as noted by Deonovic et al. (2019), the supposedly 'better fitting' generalised models typically require even greater sample sizes and can tolerate fewer attributes than saturated models. Analyses using these generalised models would likely necessitate further pooling of different national samples as individual national samples are likely too small, especially after accounting for missing data arising from the study's booklet design.

Nonetheless, I acknowledge that, for a well-designed cognitive diagnostic assessment, the use of a single saturated model may be less appropriate than multiple more generalised models, with the caveat that a single saturated model may still be preferred if the priority is to garner more easily interpretable results.

7.3.2 Data relevance

Another limitation of the analyses of this thesis is that they rely on somewhat outdated data. This is particularly pertinent because, since TIMSS 2011, there have been several changes to the ways in which data are collected in many of these international assessments, including in TIMSS. As of the most recent cycle of TIMSS, TIMSS 2023, assessment is conducted entirely digitally, rather than on paper (Mullis et al., 2021). There have also been minor changes to the topic areas within the content domains of the TIMSS assessment framework since 2011. Additionally, while TIMSS still adopts a booklet structure, the item blocks within each booklet are separated into three difficulties (easy, medium, and hard), and each student either being given an easy block followed by a medium block, or a medium block followed by a difficult block. Countries can choose to administer a higher proportion of the easy-medium or medium-high difficulty booklets, based on their prior average scores in TIMSS. This approach contrasts to the distribution of items in previous cycles, in which difficulties were somewhat evenly distributed across blocks. This means that individual students receive a more limited range of item difficulties, though at the national level, fewer students will receive booklets that are poorly calibrated to their ability levels.

These changes matter because one of the arguments of paper 3 (and the thesis as a whole) is that the structure of the data collection (as of TIMSS 2011) was a barrier to person-diagnostic modelling. It might therefore be argued that the changes to the administration of studies such as TIMSS could render this argument moot. However, my arguments about the issues associated with TIMSS data are focused on the low number of items completed by students relative to the breadth of the content of TIMSS mathematics items. This has not changed since the 2011 cycle, and so I believe these arguments are still relevant today.

Additionally, while papers 2 and 3 focused specifically on analyses of the fourth-grade mathematics assessment in TIMSS, the conclusions drawn are not specific to this age group, to TIMSS, nor just to mathematics assessment. Rather, the limitations associated with TIMSS 2011 apply more broadly to international large-scale assessments and other assessments which look to measure broad educational constructs, and I believe this thesis contributes to the wider debates about the utility of cognitive psychometric models.

7.4 Future research

One question that the thesis raises is – is there any justifiable way of utilising TIMSS data or materials for CPM research? In my view, the potential applications of TIMSS for person-diagnostic research are limited, and I question the relevance of conducting such research, given the outdated nature of the data, and of course that one of the main proposed benefits of diagnostic assessment is to be able to inform future educational guidance and intervention.

Compared to the potential of person-diagnostic modelling, I believe that item-explanatory modelling is less problematic. Although the booklet structure introduces some complexities into the analysis, the relatively low number of responses at the examinee-level is less problematic for item-focused analyses than for examinee-focused analyses. However, I do believe that the breadth of topics assessed in TIMSS mathematics relative to the total item count still presents challenges for Q-matrix construction for item-explanatory modelling. Rather than focusing on the TIMSS assessment overall, I believe the more sensible approach would be to conduct focused analyses on specific content areas that are better represented amongst the collective group of TIMSS test-items. For example, there may be greater scope to conduct such analysis on test-items relating to 2D and 3D shapes, which was a topic that was comparatively well represented among the 85 test-items included in paper 2 and paper 3. The Q-matrix used for analysis could therefore consist of more granular, content specific features that are hypothesised to relate to item difficulties.

Of course, such an analysis would not be sensible directly from the TIMSS dataset, because individual examinees have too few responses to items belonging to individual topic areas – even

if the examinees are not the focus of the analysis, stable item parameter estimation is still reliant on valid estimation of students' ability (θ) estimates. It would, however, be possible to pool relevant items from across the different booklets and conduct assessments on independent samples of students (e.g. a booklet consisting of only fractions and decimals items). TIMSS items could be supplemented by items specifically crafted to fill in gaps of the desired Q-matrix.

There are advantages to this approach over designing all new items from scratch. TIMSS test-items have already undergone thorough piloting, testing, and approval by mathematics content experts. Additionally, item-level parameters such as difficulty and discrimination can be compared to those calculated by the IEA; this can be used to help anchor items and help calibrate the difficulties of newly designed items – this would also help to facilitate in drawing comparisons between the study sample and those of the international TIMSS cohorts. It should, however, be noted that the IEA has released less and less information on TIMSS (and PIRLS) item content in more recent cycles of the study, and so future analyses may also have to rely on this increasingly outdated data.

7.5 Final remarks

While this thesis has focused on TIMSS, I believe the findings and arguments presented here apply more broadly and echo the sentiments of others who have cautioned against retrofitting cognitive psychometric models to pre-existing assessment datasets. All assessments have a purpose. Any well-designed assessment will be curated to its purpose. For assessments that include written test-items, the purpose of the assessment (i.e. what information the assessment wishes to gather) should be reflected in how test-items are created, presented, administered, and scored. The range of content covered across the collective group of test-items that form an assessment should allow for fine-grained discriminations between examinees, if that is a goal of the assessment. The development and selection of test-items should also be a carefully considered process if it wishes to draw highly specific, valid diagnostic information about examinees. I believe that a mismatch between the original purpose of an assessment, and the goals of a secondary analysis of data from that assessment will almost certainly run the risk of

yielding poor results, of which the validity should be questioned. This is regardless of the quality of the original assessment.

There is great potential for cognitive diagnostic assessments to be able to yield highly pedagogically valuable information, though it is unfortunate that the number of real-world applications of these models remains low (Ravand & Baghaei, 2020; Williamson, 2023). International assessments such as TIMSS appear, on the surface, to be an entirely sensible bridging step between simulation studies and these real-world applications. Studies such as TIMSS boast large sample sizes, have quality items that have undergone thorough piloting, and are supplemented by a wealth of additional contextual information. These studies therefore overcome some of the barriers to conducting 'real' cognitive diagnostic assessment. Retrofitting analyses also meet the calls made by those who believe that the item-level data from these international assessments are underutilised, and those who believe that ILSAs like TIMSS should be able to provide more pedagogically useful insights than they currently do (e.g. Lee et al., 2011; Holliday & Holliday, 2003; Wang, 2001; Long, 2007). But TIMSS, nor PISA or PIRLS, were ever designed with cognitive diagnosis in mind. They were not even ever designed with individual student evaluation in mind. Rather, these studies are designed to a) provide accurate estimates of the distributions of ability levels of national cohorts of students, rather than provide accurate estimates at the student level, and b) assess performance on broad subject areas (e.g. mathematics, reading, science) with lesser emphasis on domains within those subjects (e.g. geometry, informational/non-fiction reading, biology).

Despite my initial optimism about the potential applications of CPMs to international assessment data, it is, upon reflection, entirely unsurprising that TIMSS is not suitable for cognitive psychometric modelling. This thesis has demonstrated that retrofitting CPMs to TIMSS data is almost certain to produce invalid inferences, and thus questions the conclusions drawn by previous published literature in this area. Beyond TIMSS and other international assessments, this thesis also supports previous calls for greater caution and rigour when retrofitting CPMs (and in particularly, person-diagnostic CPMs) to assessment datasets that were not designed with cognitive diagnosis in mind.

8 Thesis References

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14(3*), 219-234. https://doi.org/10.1207/S15324818AME1403_2

Addey, C. (2017). Golden relics & historical standards: how the OECD is expanding global education governance through PISA for Development. *Critical Studies in Education, 58:3*, 311-325, https://doi.org/10.1080/17508487.2017.1352006

Akaike H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19,* 716-723. https://doi.org/10.1109/TAC.1974.1100705

Aldrich, C.E.A., Bookbinder, A., & Khorramdel, L. (2024). Developing the TIMSS mathematics and science achievement instruments. In: M. von Davier, B. Fishbein, & A. Kennedy (Eds.), *TIMSS 2023 Technical Report (Methods and Procedures)*, 1.1-1.23. Boston College, TIMSS & PIRLS International Study Center. Retrieved from: https://timss2023.org/methods/

Andrich, D. (2004). Controversy and the Rasch model: a characteristic of incompatible paradigms? *Medical Care, 42(1)*, 1-7. https://doi.org/10.1097/01.mlr.0000103528.48582.7c

Andrich, D., & Marais, I. (2018). Controlling bias in both constructed response and multiple-choice Items when analyzed with the dichotomous Rasch model. *Journal of Educational Measurement, 55(2)*, 281-307. https://doi.org/10.1111/jedm.12176

Andrich, D., Marais, I., & Humphry, S.M. (2016). Controlling guessing bias in the dichotomous Rasch model applied to a large-scale, vertically scaled testing program. *Educational and Psychological Measurement, 76(3)*, 412-435. https://doi.org/10.1177/0013164415594202

Baghaei, P. (2012). The application of multidimensional Rasch models in large scale assessment and validation: An empirical example. *Electronic Journal of Research in Educational Psychology, 10(1)*, 233-252. https://doi.org/10.25115/ejrep.v10i26.1493

Baghaei, P., & Kubinger, K.D. (2015). Linear logistic test modeling with R. *Practical Assessment, Research & Evaluation, 20(1).* https://doi.org/10.7275/8f33-hz58

Baker, F.B. (1993). Sensitivity of the linear logistic test model to misspecification of the weight matrix. *Applied Psychological Measurement, 17(3)*, 201-210. https://doi.org/10.1177/014662169301700301

Bezirhan, U., & von Davier, M. (2024). TIMSS achievement scaling methodology: Item response theory and population models. In: M. von Davier, B. Fishbein, & A. Kennedy (Eds.), *TIMSS 2023 Technical Report (Methods and Procedures),* 11.1-11.15. Boston College, TIMSS & PIRLS International Study Center. Retrieved from: https://timss2023.org/methods/

Birenbaum, M., Tatsuoka, C., & Xin, T. (2005). Large-scale diagnostic assessment: comparison of eighth graders' mathematics performance in the United States, Singapore and Israel. *Assessment in Education: Principles, Policy & Practice, 12(2)*, 167-181. https://doi.org/10.1080/09695940500143852

Bond, T., & Fox, C.M. (2015). *Applying the Rasch model: fundamental measurement in the human sciences (3rd ed.).* Routledge

Bokhove, C., Miyazaki, M., Komatsu, K., Chino, K., Leung, A., & Mok, I.A.C. (2019). The role of "opportunity to learn" in the geometry curriculum: a multilevel comparison of six countries. *Frontiers in Education, 4*, Article 63. https://doi.org/10.3389/feduc.2019.00063

Bulut, H.C., Bulut, O., & Arikan, S. (2022). Evaluating group differences in online reading comprehension: The impact of item properties. *International Journal of Testing, 23(1)*, 10-33. https://doi.org/10.1080/15305058.2022.2044821

Burkes, L.L. (2009). *Identifying differential item functioning related to student socioeconomic status and investigating sources related to classroom opportunities to learn.* Doctoral dissertation, University of Delaware. Retrieved from:

https://www.proquest.com/dissertations-theses/identifying-differential-item-functioning-related/docview/304879353/se-2

Casabianca, J.M., & Lewis, C. (2018). Statistical equivalence testing approaches for Mantel–Haenszel DIF analysis. *Journal of Educational and Behavioral Statistics, 43(4)*, 407-439. https://doi.org/10.3102/1076998617742410

Chen, C.M. (2013). *Examining uncertainty and misspecification of attributes in cognitive diagnostic models.* Doctoral dissertation, Columbia University. https://doi.org/10.7916/D8PC38K3

Chen, H., & Chen, J. (2015). Exploring reading comprehension skill relationships through the G-DINA model. *Educational Psychology, 36(6)*, 1049–1064. https://doi.org/10.1080/01443410.2015.1076764

Chen, H., & Chen, J. (2016). Retrofitting non-cognitive-diagnostic reading assessment under the generalized DINA model framework. *Language Assessment Quarterly, 13(3)*, 218-230. https://doi.org/10.1080/15434303.2016.1210610

Chen, Y.H. (2011). Cognitive diagnosis of mathematics performance between rural and urban students in Taiwan. *Assessment in Education: Principles, Policy & Practice, 19(2)*, 193-209. https://doi.org/10.1080/0969594X.2011.560562

Chen, Y.H., Ferron, J.M., Thompson, M.S., Gorin, J.S., & Tatsuoka, K.K. (2010). Group comparisons of mathematics performance from a cognitive diagnostic perspective. *Educational Research and Evaluation, 16(4)*, 325-343. https://doi.org/10.1080/13803611.2010.527760

Chen, Y.H., Gorin, J., Thompson, M., & Tatsuoka, K. (2006). *Verification of cognitive attributes required to solve the TIMSS-1999 mathematics items for Taiwanese students*. Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, Apr 2006). Retrieved from: https://eric.ed.gov/?id=ED491509

Chen, Y., MacDonald, G., & Leu, Y. (2011). Validating cognitive sources of mathematics item difficulty: Application of the LLTM to fraction conceptual items. *The International Journal of Educational and Psychological Assessment, 7(2*), 74-93. Retrieved from: https://www.researchgate.net/profile/Yi-Hsin-Chen-3/publication/286926672_Validating_cognitive_sources_of_mathematics_item_difficulty_Application_of_the_LLTM_to_fraction_conceptual_items/links/567029e408ae5252e6f1d43a/Validating-cognitive-sources-of-mathematics-item-difficulty-Application-of-the-LLTM-to-fraction-conceptual-items.pdf

Choi, K.M., Lee, Y.S., & Park, Y.S. (2015). What CDM can tell about what students have learned: an analysis of TIMSS eighth grade mathematics. *Eurasia Journal of Mathematics, Science and Technology Education, 11(6)*, 1563-1577. https://doi.org/10.12973/eurasia.2015.1421a

Cordero, J.M., Cristóbal, V. and Santín, D. (2018). Causal inference on education policies: a survey of empirical studies using PISA, TIMSS, and PIRLS. *Journal of Economic Surveys, 32*, 878-915. https://doi.org/10.1111/joes.12217

Daroczy, G., Wolska, M., Meurers, W.D., & Nuerk, H.C. (2015). Word problems: A review of linguistic and numerical factors contributing to their difficulty. *Frontiers in Psychology*, *6*, Article 348. https://doi.org/10.3389/fpsyg.2015.00348

De Boeck, P., Wilson, M. (2004). A framework for item response models. In: De Boeck, P., Wilson, M. (eds) *Explanatory Item Response Models. Statistics for Social Science and Public Policy*. Springer, New York, NY. https://doi.org/10.1007/978-1-4757-3990-9_1

de la Torre, J. (2009). DINA model and parameter estimation: a didactic. *Journal of Educational and Behavioral Statistics, 34(1*), 115-130. https://doi.org/10.3102/1076998607309474

de la Torre, J. (2012). The generalized DINA model framework. *Psychometrika, 76*, 179-199. https://doi.org/10.1007/s11336-011-9207-7

de la Torre, J., & Douglas, J.A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69(3)*, 333-353. https://doi.org/10.1007/BF02295640

Delafontaine, J., Chen, C., Park, J.Y., & Van den Noortgate, W. (2022). Using country-specific Q-matrices for cognitive diagnostic assessments with international large-scale data. *Large-scale Assessments in Education, 10(1)*, Article 19, https://doi.org/10.1186/s40536-022-00138-4

Deonovic, B., Chopade, P., Yudelson, M., de la Torre, J., & von Davier, A.A. (2019). Application of cognitive diagnostic models to learning and assessment systems. In: M. von Davier & Y.S. Lee (Eds.), *Handbook of diagnostic classification models: Models and model extensions, applications, software packages* (pp. 437-460). Springer.

Dogan, E., Tatsuoka, K. (2008). An international comparison using a diagnostic testing model: Turkish students' profile of mathematical skills on TIMSS-R. *Educational Studies in Mathematics, 68,* 263-272. https://doi.org/10.1007/s10649-007-9099-8

El Masri, Y.H., & Andrich, D. (2020). The trade-off between model fit, invariance, and validity: the case of PISA science assessments. *Applied Measurement in Education, 33(2)*, 174-188. https://doi.org/10.1080/08957347.2020.1732384

El Masri, Y.H., Ferrara, S., Foltz, P.W. and Baird, J. (2017). Predicting item difficulty of science national curriculum tests: the case of key stage 2 assessments. *The Curriculum Journal, 28*, 59-82. https://doi.org/10.1080/09585176.2016.1232201

Embretson, S. (1984). A general latent trait model for response processes. *Psychometrika, 49,* 175-186. https://doi.org/10.1007/BF02294171

Embretson, S.E., & Daniel, R.C. (2008). Understanding and quantifying cognitive complexity level in mathematical problem solving items. *Psychology Science, 50(3)*, Article 328. Retrieved from: https://www.psychologie-aktuell.com/fileadmin/download/PschologyScience/3-2008/02_Embretson.pdf

Ferrara, S., & Duncan, T. (2011). Comparing science achievement constructs: Targeted and achieved. *The Educational Forum, 75*, 143-156. https://doi.org/10.1080/00131725.2011.552691

Ferrara, S., Svetina, D., Skucha, S., & Davidson, A.H. (2011). Test development with performance standards and achievement growth in mind. *Educational Measurement: Issues and Practice, 30(4)*, 3-15. https://doi.org/10.1111/j.1745-3992.2011.00218.x

Fischer, G.H. (1972). *Conditional maximum-likelihood estimations of item parameters for a linear logistic test model* (Research Bulletin 9). Vienna: University of Vienna, Psychological Institute. https://doi.org/10.13140/RG.2.2.36074.72647

Fischer, G.H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*, 359-374. https://doi.org/10.1016/0001-6918(73)90003-6

Fischer, G.H. & Foreman, A.K. (1982). Some applications of logistic latent trait models with linear constraints on the parameters. *Applied Psychological Measurement, 4*, 397-416. https://doi.org/10.1177/014662168200600403

Fischman, G.E., Topper, A.M., Silova, I., Goebel, J., & Holloway, J.L. (2018). Examining the influence of international large-scale assessments on national education policies. *Journal of Education Policy, 34(4)*, 470-499. https://doi.org/10.1080/02680939.2018.1460493

Fisher-Hoch, H., & Hughes, S. (1996). *What makes mathematics exam questions difficult?* British Educational Research Association, University of Lancaster, England, 66. Retrieved from: https://www.cambridgeassessment.org.uk/Images/109643-what-makes-mathematics-exam-questions-difficult-.pdf

Fisher-Hoch, H., Hughes, S., & Bramley, T. (1997). What makes GCSE examination questions difficult? Outcomes of manipulating difficulty of GCSE questions. In: *British Educational Research Association Annual Conference* (11-14). Retrieved from:

https://www.cambridgeassessment.org.uk/Images/109646-what-makes-gcse-examination-questions-difficult-outcomes-of-manipulating-difficulty-of-gcse-questions.pdf

George, A.C., & Robitzsch, A. (2018). Focusing on interactions between content and cognition: a new perspective on gender differences in mathematical sub-competencies. *Applied Measurement in Education, 31(1)*, 79-97. https://doi.org/10.1080/08957347.2017.1391260

Gierl, M.J., & Cui, Y. (2008). Defining characteristics of diagnostic classification models and the problem of retrofitting in cognitive diagnostic assessment. *Measurement: Interdisciplinary Research and Perspectives, 6(4),* 263-268. https://doi.org/10.1080/15366360802497762

Glynn, S.M. (2012). International assessment: A Rasch model and teachers' evaluation of TIMSS science achievement items. *Journal of Research in Science Teaching, 49(10)*, 1321-1344. https://doi.org/10.1002/tea.21059

Golding, J., Richardson, M., Isaacs, T., Barnes, I., Wilkinson, D., Swensson, C. & Maris, R. (2024). *Trends in International Mathematics and Science Study (TIMSS) 2023: National report for England*. Retrieved from: https://www.gov.uk/government/publications/trends-in-international-mathematics-and-science-study-2023-england

Gorin, J.S. (2009). Diagnostic classification models: are they necessary? Commentary on Rupp and Templin (2008). *Measurement: Interdisciplinary Research and Perspectives, 7(1),* 30-33. https://doi.org/10.1080/15366360802715387

Gorur, R. (2017). Statistics and statecraft: exploring the potentials, politics and practices of international educational assessment. *Critical Studies in Education, 58(3)*, 261-265. https://doi.org/10.1080/17508487.2017.1353271

Green, K.E., & Smith, R.M. (1987). A comparison of two methods of decomposing item difficulties. *Journal of Educational Statistics, 12*, 369-381. https://doi.org/10.2307/1165055

Gu, Y., & Xu, G. (2021). Sufficient and necessary conditions for the identifiability of the Q-matrix. *Statistica Sinica, 31(1)*, 449-472. https://doi.org/10.5705/ss.202018.0410

Hambleton, R.K., & Jirka, S.J. (2006). Anchor-based methods for judgementally estimating item statistics. In: S.M. Downing & T.M. Haladyna (Eds.), *Handbook of test development* (pp. 399–420). Mahwah, NJ: Lawrence Erlbaum.

Hansen, M., Cai, L., Monroe, S., & Li, Z. (2014). Limited-information goodness-of-fit testing of diagnostic classification item response theory models. *British Journal of Mathematical and Statistical Psychology, 69, 3,* 225-252. https://doi.org/10.1111/bmsp.12074

Harrison, S., Kroehne, U., Goldhammer, F., Lüdtke, O., & Robitzsch, A. (2023). Comparing the score interpretation across modes in PISA: an investigation of how item facets affect difficulty. *Large-scale Assessments in Education, 11(1)*, Article 8. https://doi.org/10.1186/s40536-023-00157-9

Hartig, J., Frey, A., Nold, G., & Klieme, E. (2012). An application of explanatory item response modeling for model-based proficiency scaling. *Educational and Psychological Measurement, 72(4)*, 665-686. https://doi.org/10.1177/0013164411430707

Hernández-Torrano, D., Courtney, M.G.R. (2021). Modern international large-scale assessment in education: an integrative review and mapping of the literature. *Large-scale Assessments in Education, 9(17*). https://doi.org/10.1186/s40536-021-00109-1

Holliday, W.G., & Holliday, B.W. (2003). Why using international comparative math and science achievement data from TIMSS is not helpful. *The Educational Forum, 67*, 250–258. https://doi.org/10.1080/00131720309335038

Holmes, S.D., Meadows, M., Stockford, I., & He, Q. (2018). Investigating the comparability of examination difficulty using comparative judgement and Rasch modelling*. International Journal of Testing, 18(4)*, 366–391. https://doi.org/10.1080/15305058.2018.1486316

Hopfenbeck, T.N., Lenkeit, J., El Masri, Y., Cantrell, K., Ryan, J., & Baird, J. A. (2017). Lessons learned from PISA: a systematic review of peer-reviewed articles on the Programme for

International Student Assessment. *Scandinavian Journal of Educational Research, 62(3),* 333-353. https://doi.org/10.1080/00313831.2016.1258726

Hou, L., Nandakumar, R., & de la Torre, J. (2013). Differential item functioning assessment in cognitive diagnostic modeling: application of the Wald test to investigate DIF in the DINA model. *Journal of Educational Measurement, 51*, 98-125. https://doi.org/10.1111/jedm.12036

Hu, T., Yang, J., Wu, R., & Wu, X. (2021). An international comparative study of students' scientific explanation based on cognitive diagnostic assessment. *Frontiers in Psychology, 12,* Article 795497. https://doi.org/10.3389/fpsyg.2021.795497

Jang, E.E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: validity arguments for fusion model application to LanguEdge assessment. *Language Testing, 26(1)*, 31-73. https://doi.org/10.1177/0265532208097336

Janssen, R., Schepers, J., & Peres, D. (2004). Models with item and item group predictors. In: P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (189-212). New York, NY: Springer.

Janssen, R., Tuerlinckx, F., Meulders, M., & DeBoeck, P. (2000). A hierarchical IRT model for criterion-referenced measurement. *Journal of Educational and Behavioral Statistics, 25*, 285-306. https://doi.org/10.3102/10769986025003285

Junker, B.W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25(3)*, 258-272. https://doi.org/10.1177/01466210122032064

Kabiri, M., Ghazi-Tabatabaei, M., Bazargan, A., Shokoohi-Yekta, M., & Kharrazi, K. (2016). Diagnosing competency mastery in science: an application of GDM to TIMSS 2011 data. *Applied Measurement in Education, 30(1)*, 27-38. https://doi.org/10.1080/08957347.2016.1258407

Kane, M.T. (2016). Validity as the evaluation of the claims based on test scores. *Assessment in Education: Principles, Policy & Practice, 23(2)*, 309–311. https://doi.org/10.1080/0969594X.2016.1156645

Karami, H., & Nodoushan, M. A. S. (2011). Differential item functioning (DIF): current problems and future directions. *International Journal of Language Studies, 5, 3*, 133-142. Retrieved from: https://files.eric.ed.gov/fulltext/ED521872.pdf

Lan, M.C. (2014). *Exploring gender differential item functioning (DIF) on eighth grade mathematics items for the United States and Taiwan*. Doctoral dissertation, University of Washington. Retrieved from: https://www.proquest.com/dissertations-theses/exploring-gender-differential-item-functioning/docview/1622569346/se-2

Lee, Y.S., Park, Y.S., & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the U.S. national sample using the TIMSS 2007. *International Journal of Testing, 11(2)*, 144-177. https://doi.org/10.1080/15305058.2010.534571

Lee, Y.W., & Sawaki, Y. (2009). Cognitive diagnosis approaches to language assessment: An overview. *Language Assessment Quarterly, 6(3)*, 172-189. https://doi.org/10.1080/15434300902985108

Leighton, J., & Gierl, M. (Eds.). (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge University Press.

Linacre, J.M. (2002). What do infit and outfit, mean-square and standardized mean. *Rasch Measurement Transactions, 16, 2*, Article 878. Retrieved from: https://rasch.org/rmt/rmt162f.htm

Linacre J.M. (2005). Rasch dichotomous model vs. one-parameter logistic model. *Rasch Measurement Transactions, 19, 3*, Article 1032. Retrieved from: https://www.rasch.org/rmt/rmt193h.htm

Liu, X. & McKeough, A. (2005). Developmental growth in students' concept of energy: analysis of selected items from the TIMSS database. *Journal of Research in Science Teaching, 42*: 493-517. https://doi.org/10.1002/tea.20060

Long, C. (2007). *What can we learn from TIMSS 2003*. Centre for Evaluation and Assessment, University of Pretoria. Retrieved from: https://www.amesa.org.za/AMESA2007/Volumes/Vol11.pdf

Ma, L. (2014). *Validation of the item-attribute matrix in TIMSS-mathematics using multiple regression and the LSDM*. Doctoral thesis, University of Denver. Retrieved from: https://www.proquest.com/dissertations-theses/validation-item-attribute-matrix-timss/docview/1525999295/se-2

Ma, W. & de la Torre, J. (2016). A sequential cognitive diagnosis model for polytomous responses. *British Journal of Mathematical and Statistical Psychology, 69,* 253-275. https://doi.org/10.1111/bmsp.12070

Ma, W., & de la Torre, J. (2019). Category-level model selection for the sequential G-DINA model. *Journal of Educational and Behavioral Statistics, 44(1),* 45-77. https://doi.org/10.3102/1076998618792484

Ma, W. & de la Torre, J. (2020a). An empirical Q-matrix validation method for the sequential generalized DINA model. *British Journal of Mathematical and Statistical Psychology, 73*, 142-163. https://doi.org/10.1111/bmsp.12156

Ma, W. & de la Torre, J. (2020b). GDINA: An R package for cognitive diagnosis modeling. *Journal of Statistical Software*, *93(14)*, 1-26. https://doi:10.18637/jss.v093.i14

Madison, M.J., & Bradshaw, L.P. (2015). The effects of Q-matrix design on classification accuracy in the log-linear cognitive diagnosis model. *Educational and Psychological Measurement, 75(3)*, 491-511. https://doi.org/10.1177/0013164414539162

Mair, P. & Hatzinger, R. (2007). Extended Rasch modelling. The eRm package for the application of IRT models in R. *Journal of Statistical Software, 20*, 1-20. https://doi.org/10.18637/jss.v020.i09

Martin, M.O. & Mullis, I.V.S. (2012). *Methods and procedures in TIMSS and PIRLS 2011.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from: https://timssandpirls.bc.edu/methods/index.html

Martin, M.O. & Mullis, I.V.S. (2013). *TIMSS and PIRLS 2011: Relationships among reading, mathematics, and science achievement at the fourth grade - implications for early learning.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from: https://timssandpirls.bc.edu/timsspirls2011/international-database.html

Melican, G.J., Mills, C.N., & Plake, B.S. (1989). Accuracy of item performance predictions based on the Nedelsky standard setting method. *Educational and Psychological Measurement, 49,* 467-478. https://doi.org/10.1177/0013164489492020

Messick, S. (1993). Validity. In: R.L. Linn (Ed.), *Educational measurement, 3*, 3-103. New York, NY: Macmillan.

Mislevy, R.J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*, 177-196. https://doi.org/10.1007/BF02294457

Mislevy, R.J., Beaton, A.E., Kaplan, B., & Sheehan, K.M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*, 133-162. https://doi.org/10.1111/j.1745-3984.1992.tb00371.x

Müller, M. (2020). Item fit statistics for Rasch analysis: can we trust them? *Journal of Statistical Distributions and Applications, 7,* Article 5. https://doi.org/10.1186/s40488-020-00108-7

Mullis, I.V.S, Martin, M.O., Foy, P., & Arora, A. (2012). *TIMSS 2011 international results in mathematics*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston

College. Retrieved from: https://timssandpirls.bc.edu/timss2011/international-results-mathematics.html

Mullis, I.V.S., Martin, M.O., Ruddock, G.J., O'Sullivan, C.Y., Arora, A., & Erberber, E. (2005).*TIMSS 2007 assessment frameworks*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from: https://timssandpirls.bc.edu/timss2007/frameworks.html

Mullis, I.V.S., Martin, M.O., Ruddock, G.J., O'Sullivan, C.Y. & Preuschoff, C. (2009*). TIMSS 2011 assessment frameworks*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from: https://timssandpirls.bc.edu/methods/t-instrument.html

Mullis, I.V.S, Martin, M.O. & von Davier, M. (2021). *TIMSS 2023 assessment frameworks.* Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/timss2023

Nichols, P.D. (1994). A framework for developing cognitively diagnostic assessments. *Review of Educational Research, 64, 4*, 575-603. https://doi.org/10.3102/00346543064004575

Nyström, P. (2008). *Identification and analysis of text-structure and wording in TIMSS-items*. In: The 3rd IEA International Research Conference, 12. Retrieved from: https://www.iea.nl/sites/default/files/2019-04/IRC2008_Nystrom.pdf

Park, Y.S., & Lee, Y.S. (2014). An extension of the DINA model using covariates: examining factors affecting response probability and latent classification. *Applied Psychological Measurement, 38(5)*, 376-390. https://doi.org/10.1177/0146621614523830

Paulsen, J., & Valdivia, D.S. (2021). Examining cognitive diagnostic modeling in classroom assessment conditions. *The Journal of Experimental Education, 90(4)*, 916-933. https://doi.org/10.1080/00220973.2021.1891008

Pettersen, A., Nortvedt, G.A. (2018). Identifying competency demands in mathematical tasks: recognising what matters. *International Journal of Science and Mathematics Education 16*, 949-965. https://doi.org/10.1007/s10763-017-9807-5

Pollitt, A., Ahmed, A., & Crisp, V. (2007). The demands on examination syllabuses and question papers. In: P. Newton, J.A. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards*, 166-206. London: Qualifications and Curriculum Authority.

Pollitt, A., Entwistle, N., Hutchinson, C. and de Luca, C. (1985). *What makes exam questions difficult?* Scottish Academic Press, Edinburgh.

Rasch, G. (1960*). Probabilistic models for some intelligence and attainment tests*. Danish Institute for Educational Research: Copenhagen.

Ravand, H., & Baghaei, P. (2020). Diagnostic classification models: recent developments, practical issues, and prospects. *International Journal of Testing, 20(1),* 24-56. https://doi.org/10.1080/15305058.2019.1588278

Ravand, H., & Robitzsch, A. (2015). Cognitive diagnostic modeling using R. *Practical Assessment, Research, and Evaluation* 20(1): 11. https://doi.org/10.7275/5g6f-ak15

Rewer, A., & Greefrath, G. (2023). The impact of linguistic and mathematical difficulty-determining characteristics of students' performance in test-items in mathematics. In: *Thirteenth Congress of the European Society for Research in Mathematics Education (CERME13) (No. 21)*. Alfréd Rényi Institute of Mathematics; ERME. Retrieved from: https://hal.science/hal-04413560v1

Richardson, M., Isaacs, T., Barnes, I., Swensson, C., Wilkinson, D., & Golding, J. (2020). *Trends in International Maths and Science Study (TIMSS) 2019: National Report for England.* Retrieved from: https://www.gov.uk/government/publications/trends-in-international-mathematics-and-science-study-2019-england

Rosca, C.V. (2004). *What makes a science item difficult? A study of TIMSS-R items using regression and the linear logistic test model*. Doctoral dissertation, Boston College. Retrieved from: https://www.proquest.com/dissertations-theses/what-makes-science-item-difficult-study-timss-r/docview/305214920/se-2

Rupp, A.A. (2007). The answer is in the question: A guide for describing and investigating the conceptual foundations and statistical properties of cognitive psychometric models. *International Journal of Testing, 7(2)*, 95-125. https://doi.org/10.1080/15305050701193454

Rupp, A.A. (2023). Primer on diagnostic classification models (2.0 ed.). Center for Assessment. Retrieved from: https://www.nciea.org/library/primer-on-diagnostic-classificationmodels-dcms/

Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement, 74(1)*, 31-57. https://doi.org/10.1177/0013164413498257

Sandilands, D., Oliveri, M.E., Zumbo, B.D., & Ercikan, K. (2013). Investigating sources of differential item functioning in international large-scale assessments using a confirmatory approach. *International Journal of Testing, 13(2)*, 152-174. https://doi.org/10.1080/15305058.2012.690140

Schwab, C.J. (2007). *What can we learn from PISA? Investigating PISA's approach to scientific literacy*. Doctoral dissertation, University of California, Berkeley. Retrieved from: https://www.proquest.com/dissertations-theses/what-can-we-learn-pisa-investigating-pisas/docview/304899441/se-2

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics 6*, 461-464. Retrieved from: http://www.jstor.org/stable/2958889

Shaftel, J., Belton-Kocher, E., Glasnapp, D., & Poggio, J. (2006). The impact of language characteristics in mathematics test items on the performance of English language learners and students with disabilities. *Educational Assessment, 11(2)*, 105-126. https://doi.org/10.1207/s15326977ea1102_2

Sjøberg, S., & Jenkins, E. (2020). PISA: a political project and a research agenda. *Studies in Science Education, 58(1)*, 1-14. https://doi.org/10.1080/03057267.2020.1824473

Skaggs, G., Wilkins, J.L.M., & Hein, S.F. (2016). Grain size and parameter recovery with TIMSS and the general diagnostic model. *International Journal of Testing, 16(4)*, 310-330. https://doi.org/10.1080/15305058.2016.1145683

Stenner, A. (1978). Personal communication to Alastair Pollitt, as reported in Fisher-Hoch and Hughes (1996).

Stiff, J., Lenkeit, J., Hopfenbeck, T.N., Kayton, H.L., & McGrane, J.A. (2023). Research engagement in the Progress in International Reading Literacy Study: a systematic review. *Educational Research Review, 40*, Article 100547. https://doi.org/10.1016/j.edurev.2023.100547

Stiller, J., Hartmann, S., Mathesius, S., Straube, P., Tiemann, R., Nordmeier, V., Krüger, D. & Upmeier zu Belzen, A. (2016). Assessing scientific reasoning: a comprehensive evaluation of item features that affect item difficulty. *Assessment & Evaluation in Higher Education, 41(5)*, 721-732. https://doi.org/10.1080/02602938.2016.1164830

Stubbe, T.C. (2011). How do different versions of a test instrument function in a single language? A DIF analysis of the PIRLS 2006 German assessments. *Educational Research and Evaluation, 17(6)*, 465-481. https://doi.org/10.1080/13803611.2011.630560

Su, Y.L. (2013). *Cognitive diagnostic analysis using hierarchically structured skills*. Doctoral thesis, University of Iowa. Retrieved from: https://www.proquest.com/dissertations-theses/cognitive-diagnostic-analysis-using/docview/1417050106/se-2

Şen, S., & Arican, M. (2016). A diagnostic comparison of Turkish and Korean students' mathematics performances on the TIMSS 2011 assessment. *Journal of Measurement and Evaluation in Education and Psychology, 6(2)*. Retrieved from: https://dergipark.org.tr/en/pub/epod/issue/27317/287594

Şen, S., & Cohen, A.S. (2021). Sample size requirements for applying diagnostic classification models. *Frontiers in Psychology, 11*, 1-16. https://doi.org/10.3389/fpsyg.2020.621251

Tatsuoka, K.K. (1983). Rule space: an approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20(4)*, 345-354. https://doi.org/10.1111/j.1745-3984.1983.tb00212.x

Tatsuoka, K.K. (1985). A probabilistic model for diagnosing misconceptions by the pattern classification approach. *Journal of Educational Statistics, 10(1)*, 55-73. https://doi.org/10.2307/1164930

Tatsuoka, K.K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In: *Diagnostic Monitoring of Skill and Knowledge Acquisition* (eds N. Frederiksen, R. Glaser, A. Lesgold and M. Shafto), pp.453-488. Hillsdale: Erlbaum

Tatsuoka, K.K. (2009). *Cognitive assessment: An introduction to the rule space method.* Routledge.

Tatsuoka, K K., Corter, J.E., & Tatsuoka, C. (2004). Patterns of diagnosed mathematical content and process skills in TIMSS-R across a sample of 20 countries. *American Educational Research Journal, 41(4)*, 901-926. https://doi.org/10.3102/00028312041004901

Teig, N., Scherer, R., & Olsen, R. V. (2022). A systematic review of studies investigating science teaching and learning: over two decades of TIMSS and PISA. *International Journal of Science Education, 44(12)*, 2035-2058. https://doi.org/10.1080/09500693.2022.2109075

Terzi, R., & Sen, S. (2019). A nondiagnostic assessment for diagnostic purposes: Q-matrix validation and item-based model fit evaluation for the TIMSS 2011 assessment. *Sage Open, 9(1)*. https://doi.org/10.1177/2158244019832684

Toker, T. (2010). *Cognitive diagnostic assessment of TIMSS-2007 mathematics achievement items for 8th graders in Turkey*. Master's thesis, University of Denver. Retrieved from: https://www.proquest.com/dissertations-theses/cognitive-diagnostic-assessment-timss-2007/docview/759976182/se-2

Toker, T. (2016). *A comparison of latent class analysis and the mixture Rasch model: A cross-cultural comparison of 8th grade mathematics achievement in the fourth international mathematics and science study (TIMSS-2011)*. Doctoral dissertation, University of Denver. Retrieved from: https://www.proquest.com/dissertations-theses/comparison-latent-class-analysis-mixture-rasch/docview/1830757654/se-2

Toker, T., & Green, K. (2012). *An application of cognitive diagnostic assessment on TIMMS-2007 8th grade mathematics items*. Paper presented at the annual meeting of the American Educational Research Association (Vancouver, British Columbia, Canada, April 14-16, 2012). Retrieved from: https://eric.ed.gov/?id=ED543803

von Davier, M. (2018). Diagnosing diagnostic models: From von Neumann's elephant to model equivalencies and network psychometrics. *Measurement: Interdisciplinary Research and Perspectives, 16(1)*, 59-70. https://doi.org/10.1080/15366367.2018.1436827

von Davier, M. (2020). TIMSS 2019 scaling methodology: item response theory, population models, and linking across modes. In: *Methods and Procedures: TIMSS 2019 Technical Report*, 11.1-11.25. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful? In: M. von Davier & D. Hastedt (Eds.), *IERI Monograph Series: Issues and Methodologies in Large Scale Assessments (2)*, pp. 9-36). Retrieved from:

https://ierinstitute.org/fileadmin/Documents/IERI_Monograph/Volume_2/IERI_Monograph_Volume_02_Chapter_01.pdf

von Davier, M., Kennedy, A., Reynolds, K., Fishbein, B., Khorramdel, L., Aldrich, C., Bookbinder, A., Bezirhan, U., & Yin, L. (2024). *TIMSS 2023 international results in mathematics and science*. Boston College, TIMSS & PIRLS International Study Center. https://doi.org/10.6017/lse.tpisc.timss.rs6460

Wagemaker, H. (2020). Study design and evolution, and the imperatives of reliability and validity. In: *Reliability and validity of international large-scale assessment: understanding IEA's comparative studies of student achievement*. Cham: Springer International Publishing.

Walzebug, A. (2014). Is there a language-based social disadvantage in solving mathematical items*? Learning, Culture and Social Interaction, 3(2)*, 159-169. https://doi.org/10.1016/j.lcsi.2014.03.002

Wang, J. (2001). TIMSS primary and middle school data: Some technical concerns. *Educational Researcher, 30,* 17-21. https://doi.org/10.3102/0013189X030006017

Wang, W.C., & Qiu, X.L. (2019). Multilevel modeling of cognitive diagnostic assessment: the multilevel DINA example. *Applied Psychological Measurement, 43(1)*, 34-50. https://doi.org/10.1177/0146621618765713

Webb, N.L. (2007). Issues related to judging the alignment of curriculum standards and assessments*. Applied Measurement in Education, 20(1)*, 7-25. https://doi.org/10.1080/08957340709336728

Williamson, J. (2023). *Cognitive diagnostic models and how they can be useful.* Cambridge University Press & Assessment. https://doi.org/10.17863/CAM.110853

Wilson, M., De Boeck, P., & Carstensen, C.H. (2008). Explanatory item response models: A brief introduction. In: Hartig. J., Klieme, E. & Leutner, D. (Eds.), *Assessment of competencies in educational contexts*, 91-120.

Xu, G., Wang, C., & Shang, Z. (2016). On initial item selection in cognitive diagnostic computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology, 69*. https://doi.org/10.1111/bmsp.12072

Yamaguchi, K., & Okada, K. (2018). Comparison among cognitive diagnostic models for the TIMSS 2007 fourth grade mathematics assessment. *PloS One, 13(2),* https://doi.org/10.1371/journal.pone.0188691

Yao, J., Guo, Y., & Neumann, K. (2016). Towards a hypothetical learning progression of scientific explanation. *Asia-Pacific Science Education, 2(1)*, 1-17. https://doi.org/10.1186/s41029-016-0011-7

Yun, J. (2017). *Investigating structures of reading comprehension attributes at different proficiency levels: applying cognitive diagnosis models and factor analyses*. Doctoral dissertation, Florida State University. Retrieved from: https://www.proquest.com/dissertations-theses/investigating-structures-reading-comprehension/docview/2017341448/se-2

Zhan, P., Jiao, H. & Liao, D. (2018). Cognitive diagnosis modelling incorporating item response times. *British Journal of Mathematical and Statistical Psychology, 71*, 262-286. https://doi.org/10.1111/bmsp.12114

Zhan, P., Liao, M., & Bian, Y. (2018). Joint testlet cognitive diagnosis modeling for paired local item dependence in response times and response accuracy. *Frontiers in Psychology, 9*. https://doi.org/10.3389/fpsyg.2018.00607

Zhang, S., Liu, J., & Ying, Z. (2023). Statistical applications to cognitive diagnostic testing. *Annual Review of Statistics and Its Application, 10(1)*, 651-675. https://doi.org/10.1146/annurev-statistics-033021-111803

Zhang, T. (2013). *The role of reading comprehension in large-scale subject-matter assessments.* Doctoral dissertation, University of Maryland. Retrieved from:

https://www.proquest.com/dissertations-theses/role-reading-comprehension-large-scale-subject/docview/1432767383/se-2

Zhao, F. (2013). *Using noncompensatory models in cognitive diagnostic mathematics assessments: An evaluation based on empirical data*. Doctoral dissertation, University of Kansas. Retrieved from: https://www.proquest.com/dissertations-theses/using-noncompensatory-models-cognitive-diagnostic/docview/1428738807/se-2

Zhou, S., & Traynor, A. (2022). Measuring students' learning progressions in energy using cognitive diagnostic models. *Frontiers in Psychology, 13*, Article 892884. https://doi.org/10.3389/fpsyg.2022.892884

9 Tables and Figures

9.1 Tables

9.2 Figures

Appendix A – IEA approval for item access

IEA
Researching education, improving learning

Number IEA- 18-011 **(to be filled by IEA)**

PERMISSION REQUEST FORM

To be completed by anyone seeking permission to reuse, reproduce, or translate IEA material.

1. Requested IEA material

Type of requested material (select all that apply): [ ] Abstract [ ] Text excerpt [ ] Report [ ] Full article [ ] Chapter [ ] Figure/table [x] restricted use items [ ] Study instruments [ ] Other, please specify: _____

Please indicate the source of the IEA material:
Author/editor: n/a
Title: TIMSS 2011 Released Items and TIMSS 2015 Restricted Use Items
ISBN: n/a _____ Date of publication: not known _____
Description of requested IEA material (provide exact page numbers, chapter name/author, figure/table numbers, item numbers, and URL address):

I am requesting permission to use the TIMSS 2011 released items. I am also requesting access to the TIMSS 2015 restricted-use items, as well as permission for their use.

Please see https://timssandpirls.bc.edu/timss2015/international-database/assessment-items.html

NOTE: Please attach text or graphics from the IEA material you would like to use or reproduce.

3. Requestor information

First name: Jamie
Last name: Stiff
Name of institution or organization: University of Oxford
Address: Department of Education, 15 Norham Gardens
City and zip code: Oxford, OX2 6PY _____ Country: United Kingdom
Phone: +44 7807 622739 _____ Email: jamie.stiff@education.ox.ac.uk

Signature: J.Stiff Date of request: 4th January 2019

To submit your request, please sign and return this form to IEA by mail (Keizersgracht, 1016 EE Amsterdam, The Netherlands), email (secretariat@iea.nl), or fax (+31 204207136).

In signing, you agree that the permissioned material will be distributed only as part of or for use along with the original work, where the primary value does not lie with the permissioned material itself.

Please note that by signing this form you also state that you have filled out this form truthfully and to the best of your knowledge and that you have read and will comply with all conditions. Providing IEA with incorrect or incomplete information will not only invalidate permission if granted, but can also hold you liable for any damage arising from your failure to comply with these requirements. In case you have any hesitations and/or reservations regarding the information you have to fill out on this form, please contact IEA. Please allow 4-5 weeks for the processing of your permission request.

When quoting and/or citing from one or more publications and/or restricted use items from TIMSS, PIRLS and IEA for the sake of educational or research purposes, please print an acknowledgment of the source, including the year and name of that publication and/or restricted use item. Please use the following acknowledgment as an example:

SOURCE: TIMSS 2007 Assessment. Copyright © 2009 International Association for the Evaluation of Educational Achievement (IEA). Publisher: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Please note that citing without naming the source can or will constitute plagiarism for which you can and will be held accountable.

Number IEA- 18-011 **(to be filled by IEA)**

4. Permission [X] **Granted** [ ] **Denied**

Terms of agreement: Permission is granted for non-exclusive rights to reproduce the material requested above upon the terms and solely for the purpose indicated.

[X] Free [ ] Fee (payable to NL63ABNA0481961968)

Signature: _____ Date: 25 January 2019

Name: Dirk Hastedt

Title: Executive Director

Disclaimer: Please note that the website and its contents, together with all online and/or printed publications and restricted use items ('works') by TIMSS, PIRLS and other IEA studies, were created with the utmost care. However, the correctness of the information cannot be guaranteed at all times and IEA cannot and will not be held responsible or liable for any damages that may arise from the use of these resources, nor will IEA be liable for the wrongful use and/or interpretation of its works.

Please be advised that IEA cannot authorize the use of texts or items that include third-party copyrighted materials (e.g., reading passages in PIRLS, photographs, images). Users of any third-party copyrighted materials must first seek and be granted copyright permission from the owner of the content as indicated in the copyright citation line.

Please note that permission is only granted for the particular case as described in this form. Any additional use of this or any other IEA materials requires further permission. IEA copyright must be explicitly acknowledged, and the need to obtain permission for any further use of the published text/material clearly stated in the requested use/display of this material.

IEA, its proprietary assessment instruments, and studies are all the result of the choices and combination of elements by which the creator has expressed its creativity in an original matter, further to which a result has been achieved which is an intellectual creation and therefore protected as a copyright protected work, as stipulated in article 10 of the Dutch Copyright Act ("DCA") and article 2(a) of Directive 2001/29/EC regarding the harmonization of certain aspects of copyright and related rights in the information society (the "EU Copyright Directive").[1] The copyrights in these works are owned by IEA. National versions of instruments are recognized as the joint venture and shared intellectual property of the IEA and the relevant participating institutions, and should be treated accordingly.

IEA has a strict Intellectual Property Policy in place regarding third-party use of its copyright protected instruments and studies. All publications and restricted use items by TIMSS, PIRLS and other IEA studies, as well as translations thereof, are for non-commercial, educational and research purposes only. Prior permission is required when using IEA data sources for assessments or learning materials. As stated, IEA reserves the right to refuse copy deemed inappropriate or not properly sourced. IEA Intellectual Property Policy is *inter alia* included on the IEA DPC website (http://rms.iea-dpc.org/) and on the TIMSS and PIRLS website (http://timss.bc.edu/index.html), in which it is clearly stated that all accessible instruments and/or data are IEA proprietary copyright protected. Said webpages also contain links to its permission form, which should be submitted with IEA prior to any use of its materials and/or instruments.

TIMSS, PIRLS, ICCS and ICILS are registered trademarks of IEA. Use of these trademarks without permission of IEA by others may constitute trademark infringement. Furthermore, the website and its contents, together with all online and/or printed publications and restricted use items by TIMSS, PIRLS and other IEA studies are and will remain copyright of IEA.

Exploitation, distribution, redistribution, reproduction and/or transmitting in any form or by any means, including electronic or mechanical methods such as photocopying, information storage and retrieval system of these publications, restricted use items, translations thereof and/or part thereof are prohibited unless written permission has been provided by IEA.

Appendix B – Instructions and materials to reliability coders

Thank you for agreeing to help with the reliability checks of the coding of TIMSS items. This process is really helpful in identifying issues with this framework for coding the properties of mathematics test-items, and will help me improve the accuracy of my own coding.

This document briefly summarises each of the possible codes before showing a couple of example items and how I have coded them.

Each item will belong to a single content attribute, usually 1 or 2 process attributes, and any number of the skill attributes. Some of the attributes might seem quite similar – this is okay, but please feel free to ask questions if anything seems weird or incorrect!

Please also ignore any watermarks in the pictures – I do have the permission to use them!!

Content attributes

Each item can only belong to one of these five content attributes.

- **Integers (C1)**
  The item focuses on knowledge and calculations involving whole numbers that are not in a geometry or data context, and that do not involve measurement or estimation in real-world contexts. This will cover most questions that are not any of the other four content attributes.
- **Fractions and Decimals (C2)**
  The item requires an understanding a fractions and decimals. If a question includes both integer AND fraction and decimal knowledge, code it as fractions and decimals.
- **Geometry (C3)**
  The item focuses on topics such as two and three dimensional shapes, points, lines, angles, and co-ordinates.
- **Data (C4)**
  The item requires the student to interpret, complete, or perform calculations on data presented in tables and graphs.
- **Measurement and estimation (C5)**
  The item is set in a context involving measurement or estimation – this might include questions relating to time, weight, distance. These are usually problems set in real-world contexts. Problems generally set in real-world contexts but that are not directly concerned with some sort of measurement or estimation system should not be given this code.

Process attributes

Each item can belong to any number of process attributes (including none at all), though most will cover a single process attribute.

- **Translate and formulate equations (P1)**
  The item either asks the student to write down an equation, or the problem requires the student to generate an equation to solve a calculation (i.e. it is not explicitly stated / heavily suggested what the calculation should be).
- **Perform arithmetic or geometric computations (P2)**
  The item requires the student to perform some sort of higher-level computation. Include this regardless of whether the calculation is explicitly or implicitly stated. Do not assign this code if the calculation is VERY routine and would be expected for a 10-year old student to be able to recall this from memory (e.g. 3 + 6 , 5 x 7). Do include this code if the calculation involves more than two elements or more than a single step (e.g. 3 + 4 + 5 + 6) or if the calculation is more complicated (e.g. 21 x 3, or the remainder left of 19 divided

by 4). Any graphical computation (e.g. draw a shape, a bar in a graph, a line of symmetry etc.) should also have this code.

- **Make judgements or evaluate mathematical correctness (P3)**
  The item requires the student to judge whether something stated or shown is mathematically correct, or the most appropriate of a given set of choices. This is not the same as just choosing the correct answer in a multiple choice question – the student should need to perform some sort of evaluation / comparison of the given choices (i.e. "which of the following"), or the intended solution would involve checking whether each of the given answer choices is correct.
- **Reason logically, generalise, and make predictions about *verbal* information (P4)**
  The item requires the student to perform higher-level reasoning – i.e. the question does not rely on simple recall or procedural application of a common method. The question may therefore require integration of knowledge from different mathematical topics or require manipulation of information in the question in a less conventional way to solve. This code is applied to problems that require this reasoning on the text within the problem.
- **Reason logically, generalise, and make predictions about *visual* information (P5)**
  The same as P4 above, but this code is instead applied to problems that require this reasoning on visual information within the question (i.e. figures and graphs). If you think it is a combination of both, you can apply both codes.

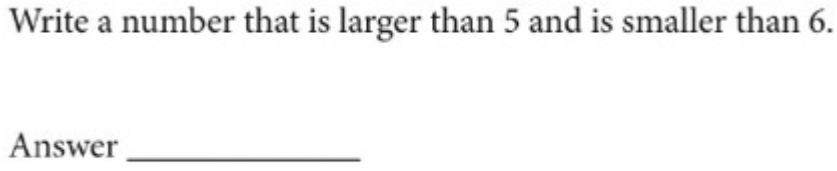
Skill attributes

- **Demonstrate number sense (S1)**
  The item tests understanding of basic number knowledge – for example, place-value (thousands, hundreds, tenths, decimal points etc.) and other number scales. Most measurement problems will also have this code.
- **Recognise and continue numerical patterns and sequences (S2)**
  The item requires the student to identify some sort of numerical pattern or sequence, and may involve continuing the pattern beyond what is shown in the question.
- **Demonstrate understanding of proportional relationships (S3)**
  The item tests understanding of proportional relationships (a bit like a pre-algebra topic) – for example, if two cakes requires three eggs, how many cakes can I make with 12 eggs).
- **Compare entities in the question (S4)**
  The item requires a comparison of two different elements in the question. This does not include comparisons between two answer options, but does include between an answer and another entity in the question itself, or two separate entities within the question (for example, comparison of data in a table and in a graph, or of a 3d shape and its net).
- **Use visual information to answer questions (S5)**
  The item includes any visual/graphical element (picture, graph etc.) that is required for the question (i.e. it would not be possible to solve the question only by reading any accompanying text. Still include this code if it can be answered through text only but some of that text is clearly a part of an image).
- **Check between options in MC items (S6)**
  The item requires some sort of checking process between the answer options given in a multiple-choice question. These are generally multiple-choice questions where there needs to be some sort of selection made and that would not be possible to answer without looking at the options – if a multiple choice question can be solved without looking at the options, do not assign this code.
- **Question requires an open-ended response (S7)**
  Assign this code for any item that does not give the student options to select from, and that could warrant any number of potential responses. Questions that do not have options A-D, but that would have a very limited set of possible answers (e.g. "What day of the week did John sell the most apples?"). Most questions with this code require either a numerical answer, require some sort of drawing, or require multiple yes/no responses.

- **Question requires verbal interpretation (S8)**
  Assign this code to an item that you feel has higher-level language demands that might be challenging for some 10 year olds. This judgement is definitely more subjective than others, but you should look for things such as a long sentence structure and more complex or technical vocabulary (including mathematical vocabulary).
- **Question is set in a novel or non-routine context (S9)**
  The question is set in a context (usually real-world but not always) that is non-routine and/or asks for a non-routine solution. Some problems may be set in simple real-world contexts but ask very routine questions (e.g. money left after a purchase) and should not be assigned this code. Again, this code is somewhat subjective.

Example items

Example item A:



Source: TIMSS 2011 Assessment. Copyright © 2013 International Association for the Evaluation of Educational Achievement (IEA). Publisher: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Example item A has been categorised as assessing the following attributes:

- C2 (Fractions and Decimals) – the student needs to understand that decimal numbers exist between integers.
- S1 (Number sense) – the student needs to identify any decimal number that lies between 5 and 6 (i.e. any values of 5.x)
- S7 – the response is open-ended.

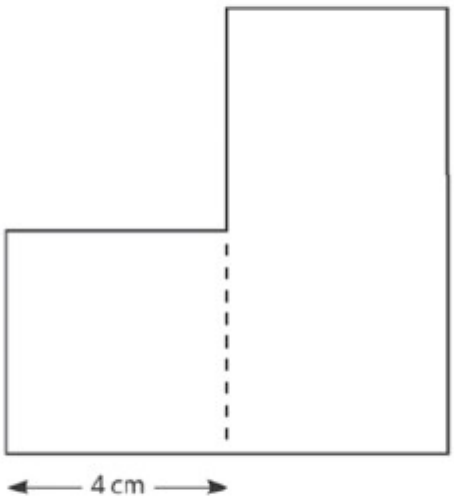
Example item B:

This shape consists of a square and a rectangle.

The width of the rectangle is the same as the width of the square.

The length of the rectangle is twice as long as its width.

Find the perimeter of the shape.



Ⓐ 28 cm

Ⓑ 32 cm

Ⓒ 36 cm

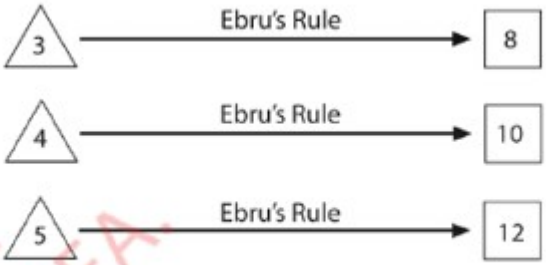
Ⓓ 40 cm

Source: TIMSS 2011 Assessment. Copyright © 2013 International Association for the Evaluation of Educational Achievement (IEA). Publisher: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Example item B has been categorised as assessing the following attributes:

- C3 (Geometry) – tests understanding of properties of squares and rectangles.
- P2 (Computations) – requires several simple numerical computations
- P4 (Verbal reasoning) – Requires verbal interpretation of the relationship between the lengths and widths of the two shapes.
- P5 (Visual reasoning) – requires visual interpretation of how the square and rectangle are combined, and how this affects the overall perimeter.
- S5 (Use visual information) – the image is required to solve the question
- S8 (Interpret verbal questions) – the question includes technical language including perimeter, length and width.

Example item C:



Ebru used a rule to get the number in the ☐ from the number in the △.
What was the rule?

Ⓐ Multiply by 1 then add 5.
Ⓑ Multiply by 2 then add 2.
Ⓒ Multiply by 3 then subtract 1.
Ⓓ Multiply by 4 then subtract 4.

Source: TIMSS 2011 Assessment. Copyright © 2013 International Association for the Evaluation of Educational Achievement (IEA). Publisher: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Example item C has been categorised as assessing the following attributes:

- C1 (Integers) – deals with integers only.
- P1 (Translate/Formulate Equations) – requires student to work out the correct computational rule.
- P2 – Repeated (simple) calculations are required to check the rules.
- P3 (Evaluate mathematical correctness) – most likely approach is to check each rule against each of the three example number sentences.
- S5 - use the image of the three number sentences.
- S6 - Most likely solution involves the student checking each of the four options until they find the rule that works for all three number sentences.

Summary of codes in example items:

Item	C1	C2	C3	C4	C5	P1	P2	P3	P4	P5	S1	S2	S3	S4	S5	S6	S7	S8	S9
A		X									X						X		
B			X				X		X	X					X			X	
C	X					X	X	X							X	X			

Appendix C – Q-matrices

This appendix provides additional details on the attributes assigned to each item in Q-matrices A – D for paper 2, as well as in Q-matrices 1 and 2 in paper 3.

Q-matrix A

Item ID	j_{A01}	j_{A02}	j_{A03}	j_{A04}	j_{A05}	j_{A06}
M031346A	0	1	0	1	0	0
M031346B	0	0	1	1	0	0
M031346C	0	0	1	1	0	0
M031379	0	0	1	1	0	0
M031380	0	0	1	1	0	0
M031313	0	1	0	1	0	0
M031083	1	0	0	0	1	0
M031071	1	0	0	0	1	0
M031185	0	0	1	1	0	0
M051305	0	1	0	1	0	0
M051091	1	0	0	1	0	0
M051001	0	0	1	1	0	0
M051007	0	0	1	1	0	0
M051203	1	0	0	1	0	0
M051601	0	1	0	1	0	0
M051064A	1	0	0	0	1	0
M051064B	0	1	0	0	1	0
M051015	0	1	0	0	1	0
M051123	1	0	0	0	1	0
M051109	1	0	0	0	0	1
M051117	0	0	1	0	0	1
M041010	1	0	0	1	0	0
M041098	0	1	0	1	0	0
M041064	0	1	0	1	0	0
M041003	1	0	0	1	0	0
M041104	1	0	0	1	0	0
M041299	1	0	0	1	0	0
M041329	1	0	0	0	1	0
M041143	1	0	0	0	1	0
M041158	0	1	0	0	1	0
M041328	0	1	0	0	1	0
M041155	0	1	0	0	1	0
M041284	0	0	1	0	1	0
M041335	1	0	0	0	0	1
M041184	0	0	1	0	0	1
M051205	1	0	0	1	0	0
M051039	0	1	0	1	0	0
M051055	0	1	0	1	0	0
M051006	0	0	1	1	0	0
M051070	0	1	0	1	0	0
M051018	0	0	1	1	0	0
M051407	0	1	0	0	1	0
M051410	1	0	0	0	1	0
M051059	0	1	0	0	1	0

M051093	0	1	0	1	0	0
M051134	0	1	0	1	0	0
M051077	0	0	1	1	0	0
M031128	0	1	0	1	0	0
M031016	0	0	1	1	0	0
M031183	0	1	0	0	1	0
M031187	1	0	0	0	1	0
M031251	0	1	0	0	1	0
M031294	0	0	1	0	1	0
M031297	0	1	0	0	0	1
M031218	0	1	0	0	0	1
M031109	1	0	0	1	0	0
M031159	0	0	1	1	0	0
M031133	0	1	0	1	0	0
M041107	0	1	0	1	0	0
M041011	0	1	0	1	0	0
M041122	1	0	0	0	0	1
M041041	0	1	0	0	1	0
M041320	0	1	0	1	0	0
M041115A	1	0	0	0	1	0
M041115B	1	0	0	0	1	0
M041160A	0	1	0	0	0	1
M041160B	0	1	0	1	0	0
M041327	1	0	0	1	0	0
M041148	1	0	0	1	0	0
M041265	1	0	0	1	0	0
M041175	1	0	0	1	0	0
M041199	0	1	0	1	0	0
M031210	0	0	1	1	0	0
M031009	1	0	0	0	1	0
M031252	0	1	0	0	1	0
M031316	0	1	0	0	1	0
M031317	1	0	0	0	1	0
M031079B	0	0	1	0	1	0
M031079C	1	0	0	0	0	1
M031004	0	0	1	0	0	1
M031043	1	0	0	1	0	0
M031325	0	1	0	1	0	0
M031088	0	1	0	1	0	0
M031093	1	0	0	1	0	0
M031155	1	0	0	1	0	0

ja01 = knowing, **ja02** = applying, **ja03** = reasoning, **ja04** = number, **ja05** = geometry, **ja06** = data

Q-matrix B

Item ID	j_{B01}	j_{B02}	j_{B03}	j_{B04}	j_{B05}	j_{B06}	j_{B07}	j_{B08}	j_{B09}	j_{B10}	j_{B11}	j_{B12}	j_{B13}
M031346A	0	1	0	0	0	1	0	0	0	0	0	0	1
M031346B	0	0	1	0	0	1	0	0	0	0	0	0	1
M031346C	0	0	1	0	0	1	0	0	0	0	0	0	1
M031379	0	0	1	0	0	1	0	0	0	0	0	0	1
M031380	0	0	1	0	0	1	0	0	0	0	0	0	1
M031313	0	1	0	0	0	1	0	0	0	0	0	0	1
M031083	1	0	0	0	0	0	0	0	0	0	1	0	0
M031071	1	0	0	0	0	0	0	0	0	0	1	0	0
M031185	0	0	1	0	0	1	0	0	0	0	0	0	0
M051305	0	1	0	0	0	0	0	1	0	0	0	0	0
M051091	1	0	0	0	0	0	1	0	0	0	0	0	0
M051001	0	0	1	0	0	1	0	0	0	0	0	0	1
M051007	0	0	1	0	0	1	0	0	0	0	0	0	0
M051203	1	0	0	0	1	0	0	0	0	0	0	0	1
M051601	0	1	0	0	0	0	0	0	1	0	0	0	1
M051064A	1	0	0	0	0	0	0	0	0	1	0	0	1
M051064B	0	1	0	0	0	0	0	0	0	1	0	0	1
M051015	0	1	0	0	0	0	0	0	0	0	1	0	1
M051123	1	0	0	0	0	0	0	0	0	0	1	0	0
M051109	1	0	0	0	0	0	0	0	0	0	0	1	1
M051117	0	0	1	0	0	0	0	0	0	0	0	1	0
M041010	1	0	0	1	0	0	0	0	0	0	0	0	0
M041098	0	1	0	0	0	1	0	0	0	0	0	0	0
M041064	0	1	0	0	0	0	1	0	0	0	0	0	1
M041003	1	0	0	1	0	0	0	0	0	0	0	0	1
M041104	1	0	0	0	0	0	0	1	0	0	0	0	1
M041299	1	0	0	0	0	0	0	1	0	0	0	0	1
M041329	1	0	0	0	0	0	0	0	0	1	0	0	0
M041143	1	0	0	0	0	0	0	0	0	0	1	0	1
M041158	0	1	0	0	0	0	0	0	0	0	1	0	0
M041328	0	1	0	0	0	0	0	0	0	0	1	0	1
M041155	0	1	0	0	0	0	0	0	0	0	1	0	0
M041284	0	0	1	0	0	0	0	0	0	0	1	0	1
M041335	1	0	0	0	0	0	0	0	0	0	0	1	0
M041184	0	0	1	0	0	0	0	0	0	0	0	1	0
M051205	1	0	0	0	1	0	0	0	0	0	0	0	1
M051039	0	1	0	0	0	1	0	0	0	0	0	0	1
M051055	0	1	0	0	0	1	0	0	0	0	0	0	1
M051006	0	0	1	0	0	1	0	0	0	0	0	0	1
M051070	0	1	0	0	0	0	0	1	0	0	0	0	0
M051018	0	0	1	0	0	0	0	0	1	0	0	0	0
M051407	0	1	0	0	0	0	0	0	0	0	1	0	0
M051410	1	0	0	0	0	0	0	0	0	0	1	0	0
M051059	0	1	0	0	0	0	0	0	0	0	1	0	1

M051093	0	0	1	0	0	0	0	0	0	0	1	0	0
M051134	0	1	0	0	0	0	0	0	0	0	0	1	1
M051077	0	1	0	0	0	0	0	0	0	0	0	1	0
M031128	1	0	0	0	1	0	0	0	0	0	0	0	1
M031016	0	0	1	0	0	1	0	0	0	0	0	0	1
M031183	0	1	0	0	0	1	0	0	0	0	0	0	1
M031187	0	1	0	1	0	0	0	0	0	0	0	0	0
M031251	0	1	0	0	0	0	0	0	1	0	0	0	0
M031294	1	0	0	0	0	0	0	0	0	0	0	1	0
M031297	0	1	0	0	0	0	0	0	0	0	1	0	1
M031218	0	1	0	0	1	0	0	0	0	0	0	0	0
M031109	1	0	0	0	0	0	0	0	0	1	0	0	0
M031159	1	0	0	0	0	0	0	0	0	0	1	0	0
M031133	0	1	0	0	0	0	0	0	0	0	0	1	1
M041107	0	1	0	1	0	0	0	0	0	0	0	0	0
M041011	1	0	0	1	0	0	0	0	0	0	0	0	0
M041122	1	0	0	1	0	0	0	0	0	0	0	0	1
M041041	1	0	0	0	1	0	0	0	0	0	0	0	0
M041320	1	0	0	0	0	0	1	0	0	0	0	0	0
M041115A	0	1	0	0	0	0	0	0	1	0	0	0	1
M041115B	0	0	1	0	0	0	0	0	1	0	0	0	1
M041160A	1	0	0	0	0	0	0	0	0	1	0	0	1
M041160B	0	1	0	0	0	0	0	0	0	1	0	0	1
M041327	0	1	0	0	0	0	0	0	0	0	1	0	1
M041148	1	0	0	0	0	0	0	0	0	0	1	0	1
M041265	0	0	1	0	0	0	0	0	0	0	1	0	0
M041175	1	0	0	0	0	0	0	0	0	0	0	1	0
M041199	0	0	1	0	0	0	0	0	0	0	0	1	0
M031210	1	0	0	0	0	0	1	0	0	0	0	0	0
M031009	0	1	0	0	0	1	0	0	0	0	0	0	1
M031252	0	1	0	0	0	0	0	0	1	0	0	0	0
M031316	1	0	0	1	0	0	0	0	0	0	0	0	1
M031317	1	0	0	1	0	0	0	0	0	0	0	0	0
M031079B	0	1	0	0	0	0	0	0	1	0	0	0	1
M031079C	0	0	1	0	0	0	0	0	1	0	0	0	1
M031004	0	1	0	0	0	0	0	0	0	1	0	0	0
M031043	0	1	0	0	0	1	0	0	0	0	0	0	0
M031325	0	1	0	0	0	0	0	0	0	1	0	0	1
M031088	0	1	0	0	0	0	0	0	0	1	0	0	0
M031093	1	0	0	0	0	0	0	0	0	0	1	0	0
M031155	0	1	0	0	0	0	0	0	0	0	0	1	0

jb01 = knowing, **jb02** = applying, **jb03** = reasoning, **jb04** = number sense / place-value, **jb05** = computations with whole numbers, **jb06** = problem-solving with whole numbers, **jb07** = equivalent fractions, **jb08** = computations with fractions and decimals, **jb09** = patterns and relationships, **jb10** = points, lines, and angles, **jb11** = 2D and 3D shapes, **jb12** = data, **jb13** = constructed response

Q-matrix C

Item ID	j_{C01}	j_{C02}	j_{C03}	j_{B04}	j_{C05}	j_{C06}	j_{C07}	j_{C08}	j_{C09}	j_{C10}	j_{C11}	j_{C12}	j_{C13}	j_{C14}	j_{C15}
M031346A	0	1	0	0	0	1	0	0	0	0	0	0	1	1	0
M031346B	0	0	1	0	0	1	0	0	0	0	0	0	1	1	0
M031346C	0	0	1	0	0	1	0	0	0	0	0	0	1	1	0
M031379	0	0	1	0	0	1	0	0	0	0	0	0	1	1	0
M031380	0	0	1	0	0	1	0	0	0	0	0	0	1	1	0
M031313	0	1	0	0	0	1	0	0	0	0	0	0	1	0	0
M031083	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0
M031071	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1
M031185	0	0	1	0	0	1	0	0	0	0	0	0	0	0	1
M051305	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0
M051091	1	0	0	0	0	0	1	0	0	0	0	0	0	0	1
M051001	0	0	1	0	0	1	0	0	0	0	0	0	1	1	0
M051007	0	0	1	0	0	1	0	0	0	0	0	0	0	1	1
M051203	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0
M051601	0	1	0	0	0	0	0	0	1	0	0	0	1	1	0
M051064A	1	0	0	0	0	0	0	0	0	1	0	0	1	1	1
M051064B	0	1	0	0	0	0	0	0	0	1	0	0	1	1	1
M051015	0	1	0	0	0	0	0	0	0	0	1	0	1	0	1
M051123	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1
M051109	1	0	0	0	0	0	0	0	0	0	0	1	1	1	0
M051117	0	0	1	0	0	0	0	0	0	0	0	1	0	1	0
M041010	1	0	0	1	0	0	0	0	0	0	0	0	0	0	1
M041098	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1
M041064	0	1	0	0	0	0	1	0	0	0	0	0	1	0	1
M041003	1	0	0	1	0	0	0	0	0	0	0	0	1	0	1
M041104	1	0	0	0	0	0	0	1	0	0	0	0	1	0	0
M041299	1	0	0	0	0	0	0	1	0	0	0	0	1	0	0
M041329	1	0	0	0	0	0	0	0	0	1	0	0	0	0	1
M041143	1	0	0	0	0	0	0	0	0	0	1	0	1	0	0
M041158	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0
M041328	0	1	0	0	0	0	0	0	0	0	1	0	1	0	1
M041155	0	1	0	0	0	0	0	0	0	0	1	0	0	0	1
M041284	0	0	1	0	0	0	0	0	0	0	1	0	1	1	0
M041335	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0
M041184	0	0	1	0	0	0	0	0	0	0	0	1	0	1	0
M051205	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0
M051039	0	1	0	0	0	1	0	0	0	0	0	0	1	0	1
M051055	0	1	0	0	0	1	0	0	0	0	0	0	1	0	1
M051006	0	0	1	0	0	1	0	0	0	0	0	0	1	1	1
M051070	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1
M051018	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0
M051407	0	1	0	0	0	0	0	0	0	0	1	0	0	0	1
M051410	1	0	0	0	0	0	0	0	0	0	1	0	0	1	1
M051059	0	1	0	0	0	0	0	0	0	0	1	0	1	0	1

M051093	0	0	1	0	0	0	0	0	0	0	1	0	0	1	1
M051134	0	1	0	0	0	0	0	0	0	0	0	1	1	1	1
M051077	0	1	0	0	0	0	0	0	0	0	0	1	0	1	0
M031128	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0
M031016	0	0	1	0	0	1	0	0	0	0	0	0	1	0	1
M031183	0	1	0	0	0	1	0	0	0	0	0	0	1	1	0
M031187	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
M031251	0	1	0	0	0	0	0	0	1	0	0	0	0	1	1
M031294	1	0	0	0	0	0	0	0	0	0	0	1	0	0	1
M031297	0	1	0	0	0	0	0	0	0	0	1	0	1	0	1
M031218	0	1	0	0	1	0	0	0	0	0	0	0	0	0	1
M031109	1	0	0	0	0	0	0	0	0	1	0	0	0	0	1
M031159	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1
M031133	0	1	0	0	0	0	0	0	0	0	0	1	1	1	0
M041107	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1
M041011	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0
M041122	1	0	0	1	0	0	0	0	0	0	0	0	1	0	0
M041041	1	0	0	0	1	0	0	0	0	0	0	0	0	0	1
M041320	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0
M041115A	0	1	0	0	0	0	0	0	1	0	0	0	1	0	1
M041115B	0	0	1	0	0	0	0	0	1	0	0	0	1	0	1
M041160A	1	0	0	0	0	0	0	0	0	1	0	0	1	0	0
M041160B	0	1	0	0	0	0	0	0	0	1	0	0	1	0	0
M041327	0	1	0	0	0	0	0	0	0	0	1	0	1	0	1
M041148	1	0	0	0	0	0	0	0	0	0	1	0	1	1	1
M041265	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0
M041175	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0
M041199	0	0	1	0	0	0	0	0	0	0	0	1	0	1	1
M031210	1	0	0	0	0	0	1	0	0	0	0	0	0	0	1
M031009	0	1	0	0	0	1	0	0	0	0	0	0	1	0	0
M031252	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0
M031316	1	0	0	1	0	0	0	0	0	0	0	0	1	0	1
M031317	1	0	0	1	0	0	0	0	0	0	0	0	0	0	1
M031079B	0	1	0	0	0	0	0	0	1	0	0	0	1	1	1
M031079C	0	0	1	0	0	0	0	0	1	0	0	0	1	1	1
M031004	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0
M031043	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1
M031325	0	1	0	0	0	0	0	0	0	1	0	0	1	0	1
M031088	0	1	0	0	0	0	0	0	0	1	0	0	0	1	1
M031093	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1
M031155	0	1	0	0	0	0	0	0	0	0	0	1	0	1	0

jc01 = knowing, **jc02** = applying, **jc03** = reasoning, **jc04** = number sense / place-value, **jc05** = computations with whole numbers, **jc06** = problem-solving with whole numbers, **jc07** = equivalent fractions, **jc08** = computations with fractions and decimals, **jc09** = patterns and relationships, **jc10** = points, lines, and angles, **jc11** = 2D and 3D shapes, **jc12** = data, **jc13** = constructed response, **jc14** = high reading demands, **jc15** = item includes mathematical/technical language

Q-matrix D

Item ID	jd01	jd02	jd03	jd04	jd05	jd06	jd07	jd08	jd09	jd10	jd11	jd12	jd13	jd14	jd15	jd16	jd17	jd18	jd19
M031346A	1	0	0	0	0	1	0	0	0	0	0	0	1	0	1	0	1	0	1
M031346B	1	0	0	0	0	1	0	0	0	0	0	0	1	0	1	0	1	0	1
M031346C	1	0	0	0	0	1	0	1	0	0	0	0	1	0	1	0	1	0	1
M031379	1	0	0	0	0	1	0	0	0	0	0	0	1	0	1	0	1	0	1
M031380	1	0	0	0	0	1	0	0	1	0	0	0	1	0	1	0	1	0	1
M031313	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0
M031083	0	0	1	0	0	0	0	0	0	1	0	0	0	1	1	1	0	0	0
M031071	0	0	1	0	0	0	1	0	0	0	0	0	0	1	1	0	0	0	0
M031185	0	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0	0	1	0
M051305	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	1	0
M051091	0	1	0	0	0	0	0	1	0	0	1	0	0	1	0	0	0	0	0
M051001	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1	1
M051007	0	0	0	0	1	0	0	0	1	1	0	0	0	0	1	0	0	1	1
M051203	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0
M051601	1	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	1
M051064A	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	1
M051064B	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	1
M051015	0	0	1	0	0	0	1	1	0	0	0	0	0	0	1	0	1	1	0
M051123	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0
M051109	0	0	0	1	0	0	0	1	0	0	0	0	1	0	1	0	1	0	1
M051117	0	0	0	1	0	0	1	0	0	1	0	0	0	1	1	0	0	1	1
M041010	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0
M041098	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0
M041064	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
M041003	1	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0	0	1	1
M041104	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0
M041299	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
M041329	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0
M041143	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0
M041158	1	0	0	0	0	0	1	0	0	1	0	0	0	0	1	0	0	1	1
M041328	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0
M041155	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	1
M041284	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0
M041335	0	0	0	1	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0
M041184	0	0	0	1	0	0	0	0	0	1	0	0	0	1	1	1	0	1	0
M051205	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0
M051039	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0
M051055	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	1	0
M051006	1	0	0	0	0	1	1	0	1	1	0	0	0	0	1	0	1	0	1
M051070	0	1	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0	1	0
M051018	1	0	0	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0	1
M051407	0	0	1	0	0	0	0	1	0	1	0	0	0	0	1	1	0	0	1
M051410	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	1	0	1	0
M051059	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0
M051093	0	0	1	0	0	0	1	0	1	1	0	0	0	0	1	0	0	1	0

M051134	0	0	0	1	0	0	0	0	0	1	0	0	0	1	1	0	1	1	1
M051077	0	0	0	1	0	0	0	0	0	1	0	0	0	1	1	0	0	1	1
M031128	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0
M031016	1	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1	1	1
M031183	0	1	0	0	0	1	1	0	0	0	0	0	0	0	1	0	1	1	1
M031187	1	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	1	1
M031251	1	0	0	0	0	1	1	0	0	0	0	0	0	0	1	1	0	0	0
M031294	0	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0
M031297	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0	1	1	0
M031218	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	1	1
M031109	0	0	1	0	0	0	0	1	0	0	0	0	0	1	1	1	0	0	0
M031159	0	0	1	0	0	0	0	1	0	0	0	1	0	0	1	1	0	0	0
M031133	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	0	1	1	1
M041107	1	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	1	1
M041011	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0
M041122	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1	0	0
M041041	1	0	0	0	0	0	0	1	0	0	0	0	0	1	1	1	0	0	0
M041320	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0
M041115A	1	0	0	0	0	1	0	0	1	0	0	1	0	0	1	0	1	0	0
M041115B	1	0	0	0	0	1	1	0	0	0	0	1	0	0	1	0	1	0	0
M041160A	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0
M041160B	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0
M041327	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0
M041148	0	0	1	0	0	0	0	1	0	0	0	0	0	1	1	0	1	1	0
M041265	0	0	1	0	0	0	0	1	0	0	0	0	0	1	1	1	0	0	0
M041175	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
M041199	0	0	0	1	0	0	0	1	0	0	0	0	0	1	1	0	0	0	0
M031210	0	1	0	0	0	0	1	1	0	0	1	0	0	1	0	1	0	0	0
M031009	1	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	1	1	1
M031252	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0
M031316	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0
M031317	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0
M031079B	1	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	1	0	0
M031079C	1	0	0	0	0	0	0	0	1	0	0	1	0	1	1	0	1	0	0
M031004	0	0	0	0	1	0	1	1	0	1	0	0	0	0	1	0	0	1	1
M031043	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1	1
M031325	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0
M031088	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0	1	1
M031093	0	0	1	0	0	0	0	1	0	0	0	0	0	1	1	1	0	0	0
M031155	0	0	0	1	0	0	1	0	0	0	0	0	0	1	1	0	0	1	0

jd01 = integers, **jd02** = fractions and decimals, **jd03** = geometry, **jd04** = data, **jd05** = measurement and estimation, **jd06** = translate and formulate equations, **jd07** = perform computations in arithmetic / geometry, **jd08** = make judgements or evaluate mathematical correctness, **jd09** = logical reasoning in written contexts, **jd10** = logical reasoning in visual contexts, **jd11** = number sense / place-value, **jd12** = patterns and sequences, **jd13** = proportional relationships, **jd14** = compare entities, **jd15** = item includes visual elements, **jd16** = check and evaluate MC options, **jd17** = open-ended response, **jd18** = higher-level verbal interpretation, **jd19** = novel context

Table 30: Q-matrix 1

Block	Item ID	Attributes								
		Number content			Geometry content			Data content		
		Knowing	Applying	Reasoning	Knowing	Applying	Reasoning	Knowing	Applying	Reasoning
		j_{101}	j_{102}	j_{103}	j_{104}	j_{105}	j_{106}	j_{107}	j_{108}	j_{109}
M01	M031346A	0	1	0	0	0	0	0	0	0
M01	M031346B	0	0	1	0	0	0	0	0	0
M01	M031346C	0	0	1	0	0	0	0	0	0
M01	M031379	0	0	1	0	0	0	0	0	0
M01	M031380	0	0	1	0	0	0	0	0	0
M01	M031313	0	1	0	0	0	0	0	0	0
M01	M031083	0	0	0	1	0	0	0	0	0
M01	M031071	0	0	0	1	0	0	0	0	0
M01	M031185	0	0	1	0	0	0	0	0	0
M02	M051305	0	1	0	0	0	0	0	0	0
M02	M051091	1	0	0	0	0	0	0	0	0
M02	M051001	0	0	1	0	0	0	0	0	0
M02	M051007	0	0	1	0	0	0	0	0	0
M02	M051203	1	0	0	0	0	0	0	0	0
M02	M051601	0	1	0	0	0	0	0	0	0
M02	M051064A	0	0	0	1	0	0	0	0	0
M02	M051064B	0	0	0	0	1	0	0	0	0
M02	M051015	0	0	0	0	1	0	0	0	0
M02	M051123	0	0	0	1	0	0	0	0	0
M02	M051109	0	0	0	0	0	0	1	0	0
M02	M051117	0	0	0	0	0	0	0	0	1
M03	M041010	1	0	0	0	0	0	0	0	0
M03	M041098	0	1	0	0	0	0	0	0	0
M03	M041064	0	1	0	0	0	0	0	0	0
M03	M041003	1	0	0	0	0	0	0	0	0
M03	M041104	1	0	0	0	0	0	0	0	0
M03	M041299	1	0	0	0	0	0	0	0	0

M03	M041329	0	0	0	1	0	0	0	0	0
M03	M041143	0	0	0	1	0	0	0	0	0
M03	M041158	0	0	0	0	1	0	0	0	0
M03	M041328	0	0	0	0	1	0	0	0	0
M03	M041155	0	0	0	0	1	0	0	0	0
M03	M041284	0	0	0	0	0	1	0	0	0
M03	M041335	0	0	0	0	0	0	1	0	0
M03	M041184	0	0	0	0	0	0	0	0	1
M04	M051205	1	0	0	0	0	0	0	0	0
M04	M051039	0	1	0	0	0	0	0	0	0
M04	M051055	0	1	0	0	0	0	0	0	0
M04	M051006	0	0	1	0	0	0	0	0	0
M04	M051070	0	1	0	0	0	0	0	0	0
M04	M051018	0	0	1	0	0	0	0	0	0
M04	M051407	0	0	0	0	1	0	0	0	0
M04	M051410	0	0	0	1	0	0	0	0	0
M04	M051059	0	0	0	0	1	0	0	0	0
M04	M051093	0	0	0	0	0	1	0	0	0
M04	M051134	0	0	0	0	0	0	0	1	0
M04	M051077	0	0	0	0	0	0	0	1	0
M05	M031128	1	0	0	0	0	0	0	0	0
M05	M031016	0	0	1	0	0	0	0	0	0
M05	M031183	0	1	0	0	0	0	0	0	0
M05	M031187	0	1	0	0	0	0	0	0	0
M05	M031251	0	1	0	0	0	0	0	0	0
M05	M031294	0	0	0	0	0	0	1	0	0
M05	M031297	0	0	0	0	1	0	0	0	0
M05	M031218	0	1	0	0	0	0	0	0	0
M05	M031109	0	0	0	1	0	0	0	0	0
M05	M031159	0	0	0	1	0	0	0	0	0
M05	M031133	0	0	0	0	0	0	0	1	0
M06	M041107	0	1	0	0	0	0	0	0	0
M06	M041011	1	0	0	0	0	0	0	0	0
M06	M041122	1	0	0	0	0	0	0	0	0

M06	M041041	1	0	0	0	0	0	0	0	0
M06	M041320	1	0	0	0	0	0	0	0	0
M06	M041115A	0	1	0	0	0	0	0	0	0
M06	M041115B	0	0	1	0	0	0	0	0	0
M06	M041160A	0	0	0	1	0	0	0	0	0
M06	M041160B	0	0	0	0	1	0	0	0	0
M06	M041327	0	0	0	0	1	0	0	0	0
M06	M041148	0	0	0	1	0	0	0	0	0
M06	M041265	0	0	0	0	0	1	0	0	0
M06	M041175	0	0	0	0	0	0	1	0	0
M06	M041199	0	0	0	0	0	0	0	0	1
M07	M031210	1	0	0	0	0	0	0	0	0
M07	M031009	0	1	0	0	0	0	0	0	0
M07	M031252	0	1	0	0	0	0	0	0	0
M07	M031316	1	0	0	0	0	0	0	0	0
M07	M031317	1	0	0	0	0	0	0	0	0
M07	M031079B	0	1	0	0	0	0	0	0	0
M07	M031079C	0	0	1	0	0	0	0	0	0
M07	M031004	0	0	0	0	1	0	0	0	0
M07	M031043	0	1	0	0	0	0	0	0	0
M07	M031325	0	0	0	0	1	0	0	0	0
M07	M031088	0	0	0	0	1	0	0	0	0
M07	M031093	0	0	0	1	0	0	0	0	0
M07	M031155	0	0	0	0	0	0	0	1	0

Table 31: Q-matrix 2

Block	Item ID	Attributes											
		Content attributes				Process attributes			Skill attributes				
		j_{201}	j_{202}	j_{203}	j_{204}	j_{205}	j_{206}	j_{207}	j_{208}	j_{209}	j_{210}	j_{211}	j_{212}
M01	M031346A	1	0	0	0	0	0	0	0	0	1	0	0
M01	M031346B	1	0	0	0	0	0	0	0	0	1	0	0
M01	M031346C	1	0	0	0	0	1	0	0	0	1	0	0
M01	M031379	1	0	0	0	0	0	0	0	0	1	0	0
M01	M031380	1	0	0	0	0	0	1	0	0	1	0	0
M01	M031313	1	0	0	0	1	0	0	0	0	0	0	0
M01	M031083	0	1	0	0	0	0	1	0	0	0	1	1
M01	M031071	0	1	0	0	1	0	0	0	0	0	1	0
M01	M031185	0	0	0	1	0	0	0	0	0	1	0	0
M02	M051305	0	0	0	1	1	0	0	0	0	0	0	0
M02	M051091	1	0	0	0	0	1	0	1	0	0	1	0
M02	M051001	1	0	0	0	0	0	1	0	0	0	0	0
M02	M051007	0	0	0	1	0	0	1	0	0	0	0	0
M02	M051203	1	0	0	0	1	0	0	0	0	0	0	0
M02	M051601	1	0	0	0	0	0	1	0	1	0	0	0
M02	M051064A	0	1	0	0	0	0	0	0	0	0	1	0
M02	M051064B	0	1	0	0	0	0	0	0	0	0	1	0
M02	M051015	0	1	0	0	1	1	0	0	0	0	0	0
M02	M051123	0	1	0	0	0	1	0	0	0	0	0	0
M02	M051109	0	0	1	0	0	1	0	0	0	1	0	0
M02	M051117	0	0	1	0	1	0	1	0	0	0	1	0
M03	M041010	1	0	0	0	0	0	0	1	0	0	0	1
M03	M041098	1	0	0	0	1	0	0	0	0	0	0	0
M03	M041064	1	0	0	0	0	0	0	0	0	0	0	0
M03	M041003	1	0	0	0	0	0	1	1	0	0	1	0
M03	M041104	1	0	0	0	0	0	0	1	0	0	0	0
M03	M041299	1	0	0	0	1	0	0	0	0	0	0	0
M03	M041329	0	1	0	0	0	0	0	0	0	0	0	1
M03	M041143	0	1	0	0	0	0	0	0	0	0	0	0

M03	M041158	1	0	0	0	1	0	1	0	0	0	0	0
M03	M041328	0	1	0	0	0	0	0	0	0	0	0	0
M03	M041155	0	1	0	0	1	0	0	0	0	0	0	0
M03	M041284	0	1	0	0	0	0	1	0	0	0	0	1
M03	M041335	0	0	1	0	0	0	1	0	0	0	1	0
M03	M041184	0	0	1	0	0	0	1	0	0	0	1	1
M04	M051205	1	0	0	0	1	0	0	0	0	0	0	0
M04	M051039	1	0	0	0	1	0	0	0	0	0	0	0
M04	M051055	0	0	0	1	0	0	0	1	0	0	0	0
M04	M051006	1	0	0	0	1	0	1	0	0	0	0	0
M04	M051070	1	0	0	0	1	0	0	1	0	0	0	0
M04	M051018	1	0	0	0	1	0	0	0	1	0	0	0
M04	M051407	0	1	0	0	0	1	1	0	0	0	0	1
M04	M051410	0	1	0	0	0	1	0	0	0	0	0	1
M04	M051059	0	1	0	0	0	0	0	0	0	0	0	0
M04	M051093	0	1	0	0	1	0	1	0	0	0	0	0
M04	M051134	0	0	1	0	0	0	1	0	0	0	1	0
M04	M051077	0	0	1	0	0	0	1	0	0	0	1	0
M05	M031128	1	0	0	0	1	0	0	0	0	0	0	0
M05	M031016	1	0	0	0	0	0	1	1	0	0	0	0
M05	M031183	1	0	0	0	1	0	0	0	0	0	0	0
M05	M031187	1	0	0	0	0	1	0	0	0	0	0	1
M05	M031251	1	0	0	0	1	0	0	0	0	0	0	1
M05	M031294	0	0	0	1	0	0	0	1	0	0	0	0
M05	M031297	0	1	0	0	1	0	0	0	0	0	0	0
M05	M031218	1	0	0	0	0	0	0	0	0	0	0	1
M05	M031109	0	1	0	0	0	1	0	0	0	0	1	1
M05	M031159	0	1	0	0	0	1	0	0	1	0	0	1
M05	M031133	0	0	1	0	0	0	0	0	0	0	1	0
M06	M041107	1	0	0	0	0	1	0	0	0	0	0	1
M06	M041011	1	0	0	0	0	0	0	1	0	0	0	1
M06	M041122	1	0	0	0	0	1	0	0	0	0	0	1
M06	M041041	1	0	0	0	0	1	0	0	0	0	1	1
M06	M041320	1	0	0	0	0	1	0	0	0	0	0	1

M06	M041115A	1	0	0	0	0	0	1	0	1	0	0	0
M06	M041115B	1	0	0	0	1	0	0	0	1	0	0	0
M06	M041160A	0	1	0	0	0	0	0	0	0	0	0	0
M06	M041160B	0	1	0	0	0	0	0	0	0	0	0	0
M06	M041327	0	1	0	0	0	0	0	0	0	0	0	0
M06	M041148	0	1	0	0	0	1	0	0	0	0	1	0
M06	M041265	0	1	0	0	0	1	0	0	0	0	1	1
M06	M041175	0	0	1	0	0	0	0	0	0	0	1	0
M06	M041199	0	0	1	0	0	1	0	0	0	0	1	0
M07	M031210	1	0	0	0	1	1	0	1	0	0	1	1
M07	M031009	1	0	0	0	1	0	1	0	0	0	0	0
M07	M031252	1	0	0	0	0	0	0	0	1	0	0	1
M07	M031316	1	0	0	0	0	0	0	1	0	0	0	0
M07	M031317	1	0	0	0	0	0	0	1	0	0	0	0
M07	M031079B	1	0	0	0	0	0	0	0	1	0	1	0
M07	M031079C	1	0	0	0	0	0	1	0	1	0	1	0
M07	M031004	0	0	0	1	1	1	1	0	0	0	0	0
M07	M031043	0	0	0	1	1	0	0	0	0	0	0	0
M07	M031325	0	1	0	0	1	0	0	0	0	0	0	0
M07	M031088	0	1	0	0	1	0	0	0	0	0	0	0
M07	M031093	0	1	0	0	0	1	0	0	0	0	1	1
M07	M031155	0	0	1	0	1	0	0	0	0	0	1	0

Appendix D – Item-fit statistics (paper 2)

Table 32 reports the outfit statistical residuals as calculated in paper 2 for identifying misfitting items. Outfit residuals greater than 1.3 are highlighted.

Table 32: Outfit residuals (Paper 2)

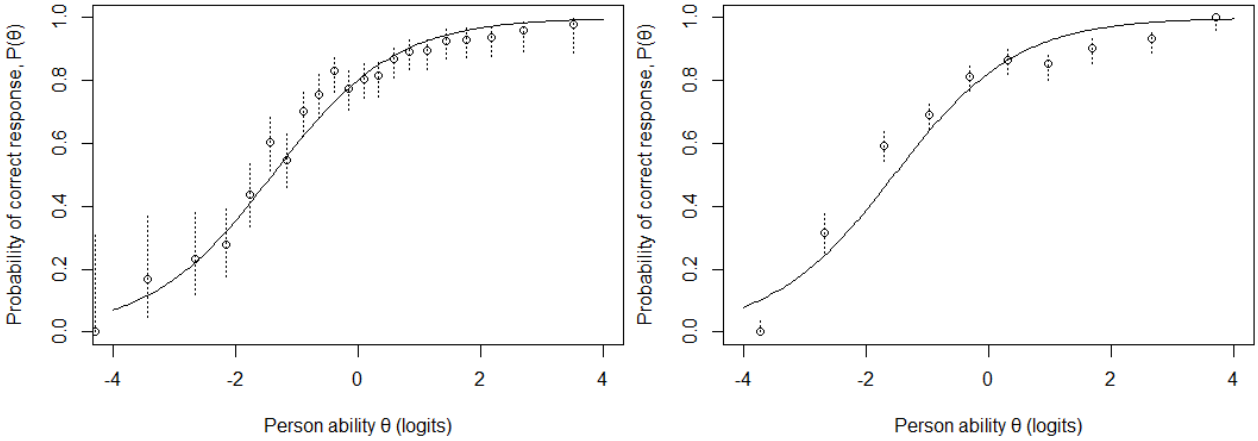
Item ID	Outfit residual A	Outfit residual B	Outfit > 1.3 ?
M031346A	0.791	0.660	
M031346B	0.793	0.634	
M031346C	0.700	0.623	
M031379	1.070	1.041	
M031380	1.103	1.272	
M031313	1.283	1.790	Yes
M031083	1.025	1.156	
M031071	1.195	1.392	Yes
M031185	0.854	1.066	
M051305	1.060	1.057	
M051091	0.803	0.827	
M051001	0.906	0.938	
M051007	1.507	1.708	Yes
M051203	1.090	1.099	
M051601	0.951	0.895	
M051064A	1.062	1.054	
M051064B	1.697	1.460	Yes
M051015	0.947	0.966	
M051123	1.284	1.139	
M051109	0.768	0.760	
M051117	1.171	1.226	
M041010	0.946	0.902	
M041098	1.122	1.067	
M041064	0.864	0.777	
M041003	1.054	0.915	
M041104	0.774	0.742	
M041299	0.705	0.726	
M041329	1.051	1.073	
M041143	1.217	1.181	
M041158	1.107	1.195	
M041328	0.913	0.931	
M041155	1.103	1.077	
M041284	0.789	0.718	
M041335	1.036	0.970	
M041184	1.228	1.195	
M051205	1.058	0.981	
M051039	0.803	0.888	
M051055	0.766	0.831	
M051006	0.891	0.942	
M051070	1.185	1.233	
M051018	1.309	1.329	Yes

M051407	1.095	1.051	
M051410	1.055	1.085	
M051059	1.019	1.155	
M051093	1.289	1.240	
M051134	0.763	0.771	
M051077	0.907	0.956	
M031128	1.256	1.113	
M031016	0.766	0.789	
M031183	0.834	0.787	
M031187	1.050	1.124	
M031251	1.019	1.024	
M031294	0.983	0.809	
M031297	1.082	0.976	
M031218	0.912	0.890	
M031109	1.205	1.031	
M031159	1.075	0.936	
M031133	**1.399**	1.274	**Yes**
M041107	0.765	0.888	
M041011	0.922	0.907	
M041122	0.840	0.819	
M041041	1.205	1.183	
M041320	0.747	0.739	
M041115A	1.081	1.047	
M041115B	0.806	0.873	
M041160A	1.243	1.163	
M041160B	0.948	0.855	
M041327	0.923	1.047	
M041148	1.110	1.120	
M041265	**1.432**	**1.350**	**Yes**
M041175	1.054	0.964	
M041199	0.738	0.821	
M031210	1.071	0.947	
M031009	0.921	0.878	
M031252	0.959	1.007	
M031316	1.001	0.959	
M031317	0.977	1.081	
M031079B	0.887	0.833	
M031079C	1.015	1.159	
M031004	1.266	**1.431**	**Yes**
M031043	0.957	1.002	
M031325	0.882	0.856	
M031088	1.104	1.064	
M031093	**1.401**	**1.347**	**Yes**
M031155	0.999	1.066	

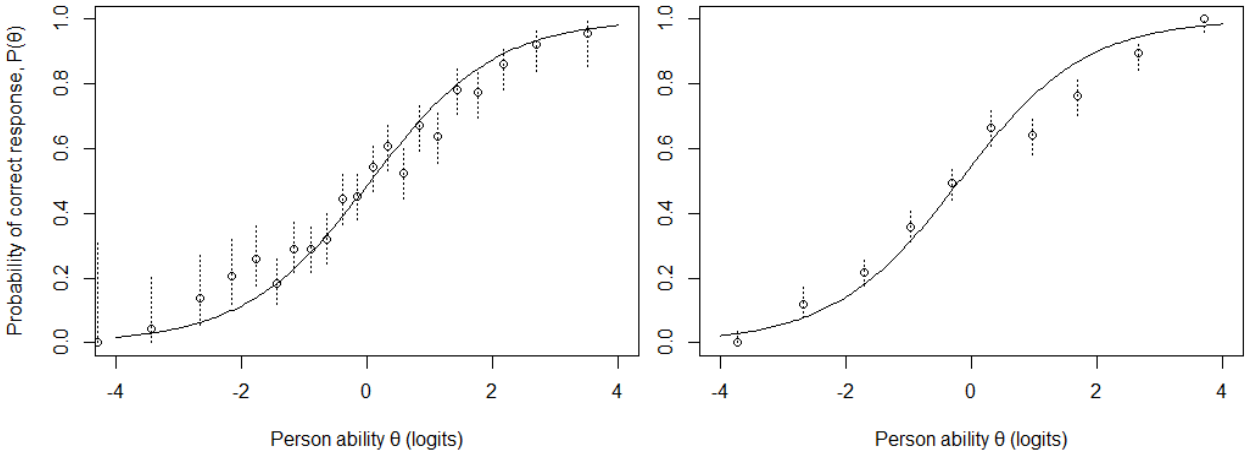
For each item with at least one outfit residual greater than 1.3, two item characteristic curves were plotted with an overlay of the probabilities of success for examinees with each raw score total for the two associated booklets. No items were removed from the analysis presented in

Paper 2, though Item 13 was identified as potentially be problematic given the relatively high outfit residuals and a clear pattern of deviance from the 1PL shared across both booklets.

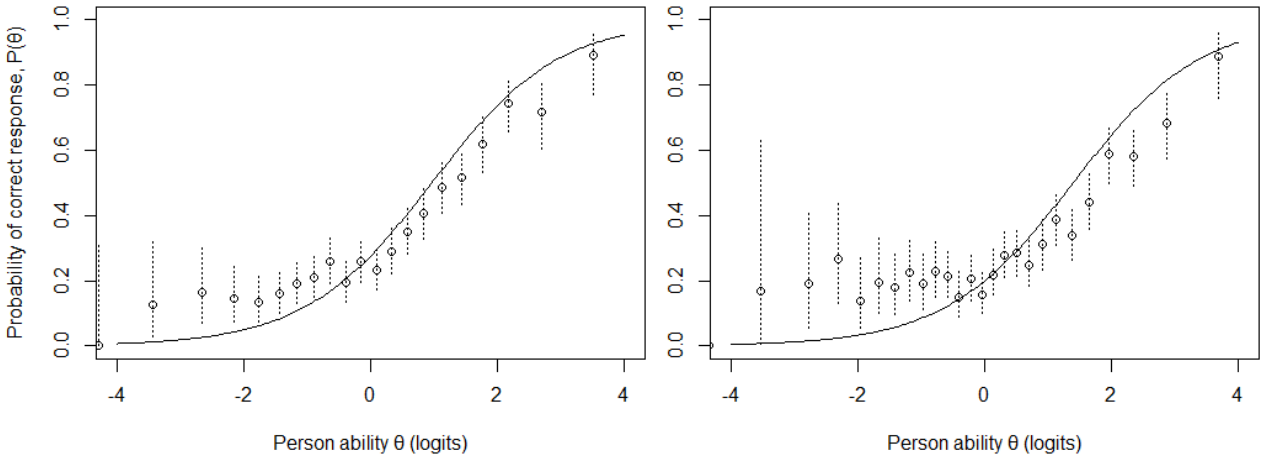
Item 6, M031313



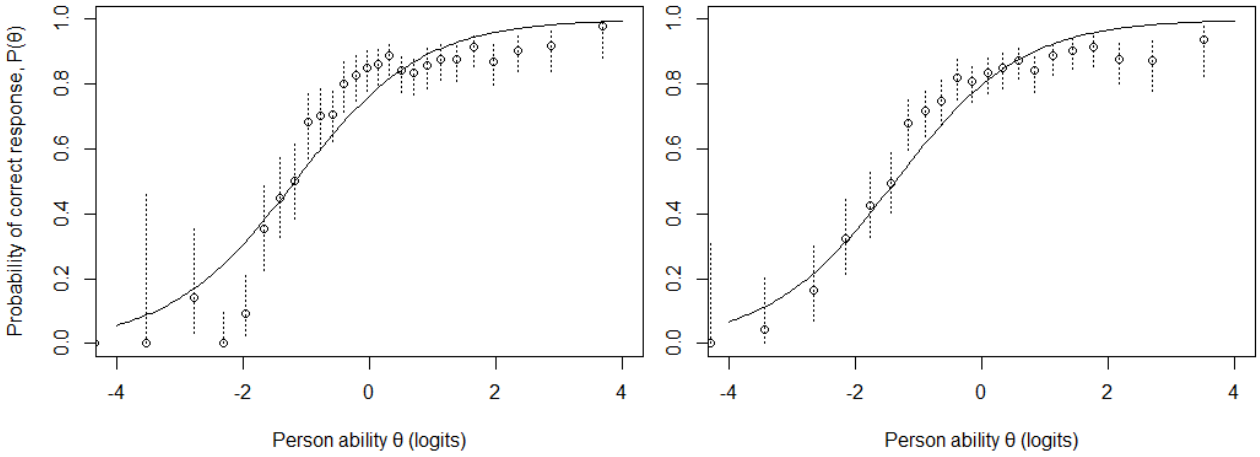
Item 8, M031071



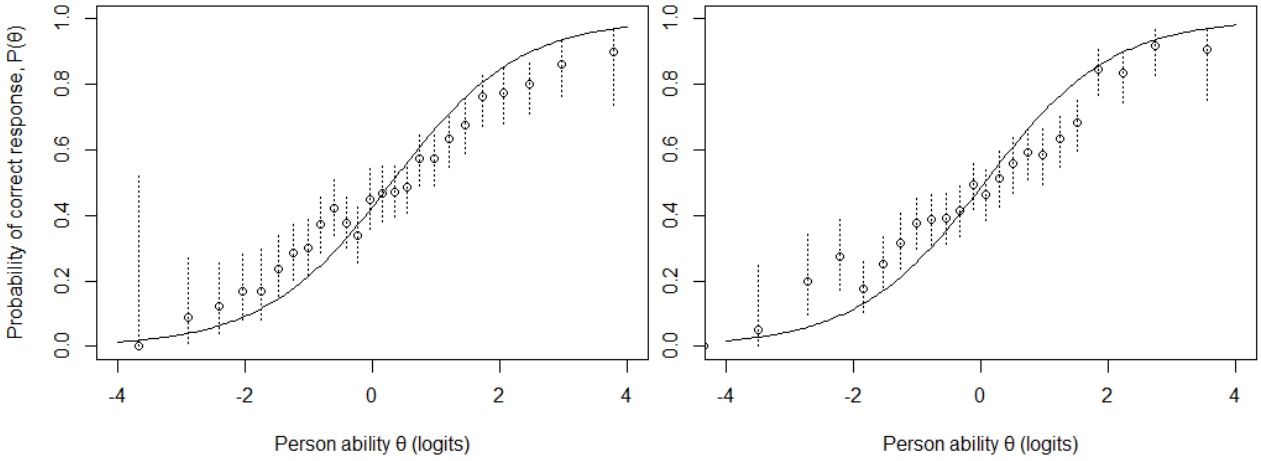
Item 13, M051007



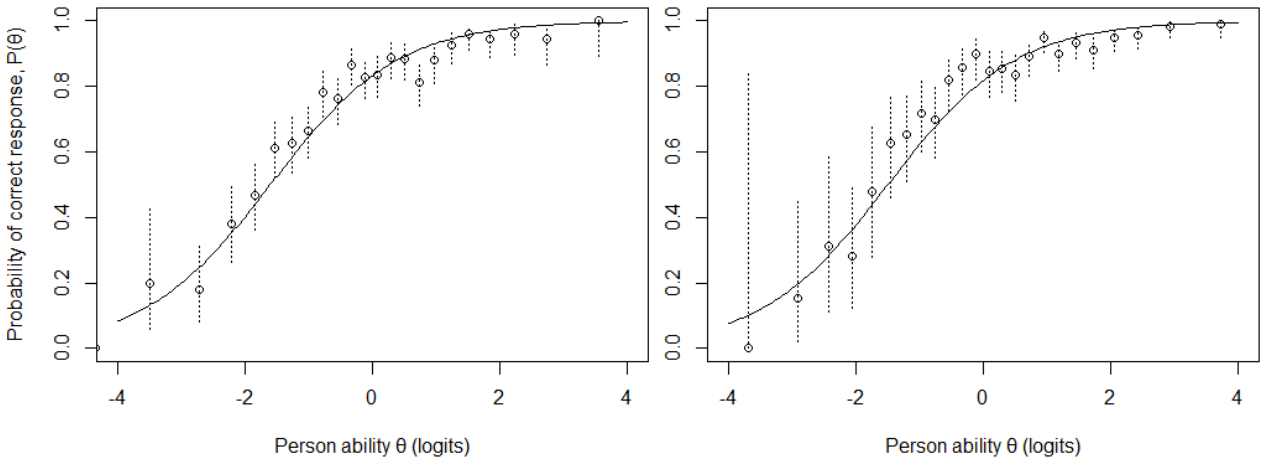
Item 17, M051064B



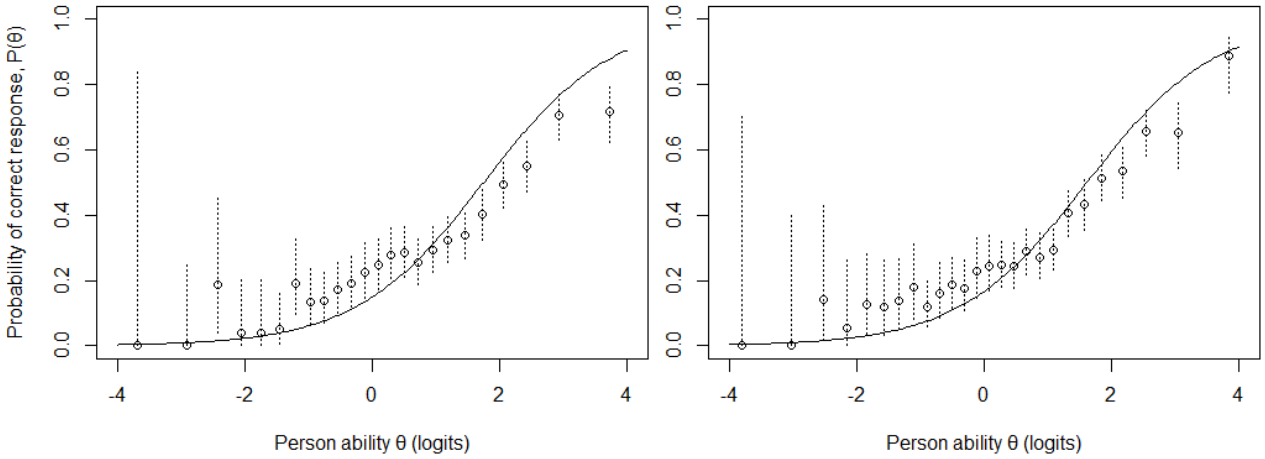
Item 41, M051018



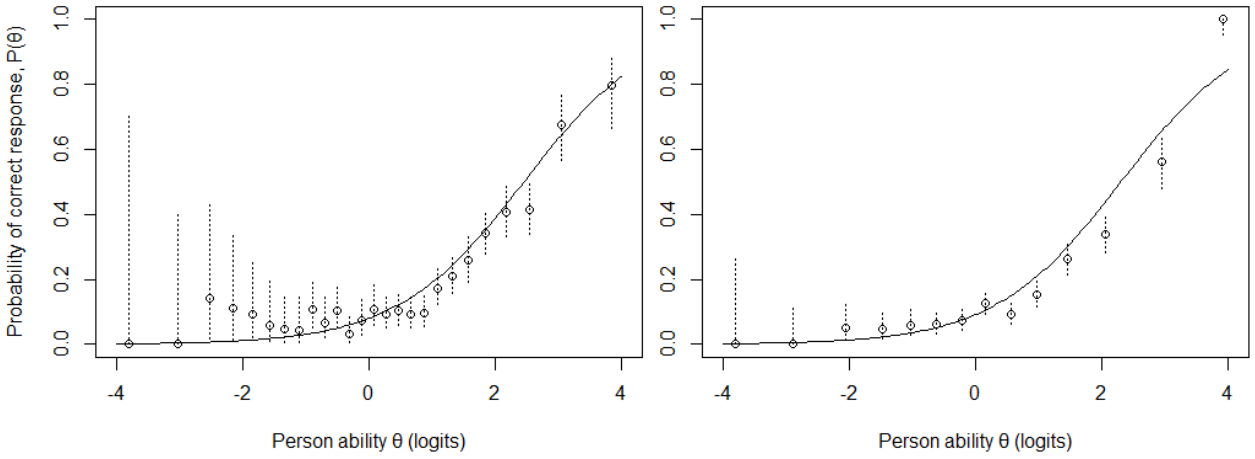
Item 58, M031133



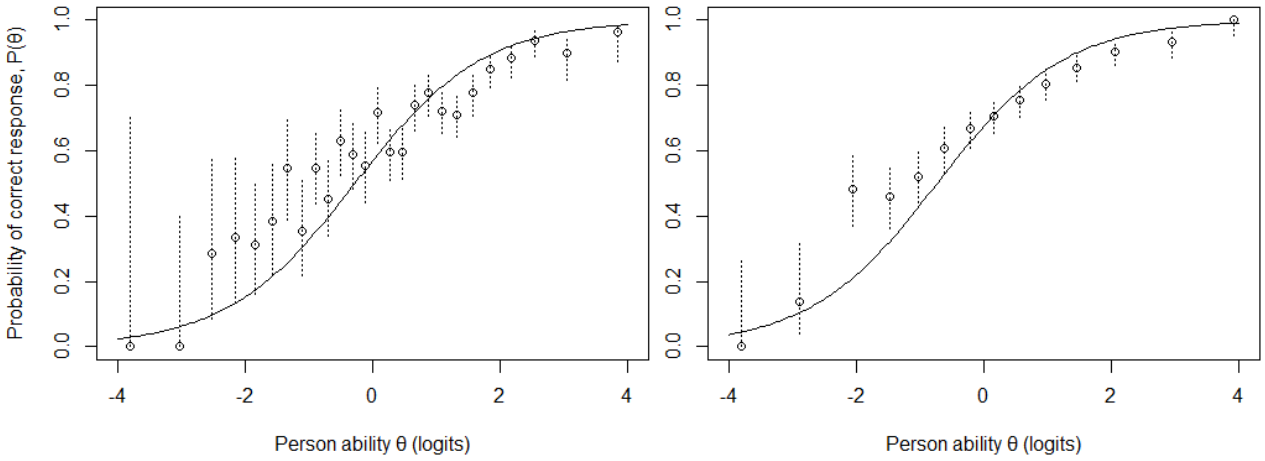
Item 70, M041265



Item 80, M031004



Item 84, M031093



Appendix E – Q-matrix 2 model fit

The original intention in paper 3 was to model attribute mastery using the best fitting Q-matrix from paper 2. This was Q-matrix D. However, Q-matrix D consisted of 19 attributes. By contrast, it has previously been reported that the DINA model and other person-diagnostic models perform best when there around five or fewer attributes. An analysis of 19 attributes is almost computationally demanding. I consequently decided it would be more appropriate to use a reduced model consisting of 12 attributes. The full process and rationale for choosing / excluding attributes was presented in section 3.4.1.

To assess the potential impact of attribute reduction on model fit, I assessed the fit of Q-matrix 2 relative to Q-matrix D. Table 33 presents the eta difficulty estimates associated with the 12 attributes of Q-matrix 2. As expected from the results of Q-matrix D, all 12 of these attributes had significant unique contributions to item difficulties.

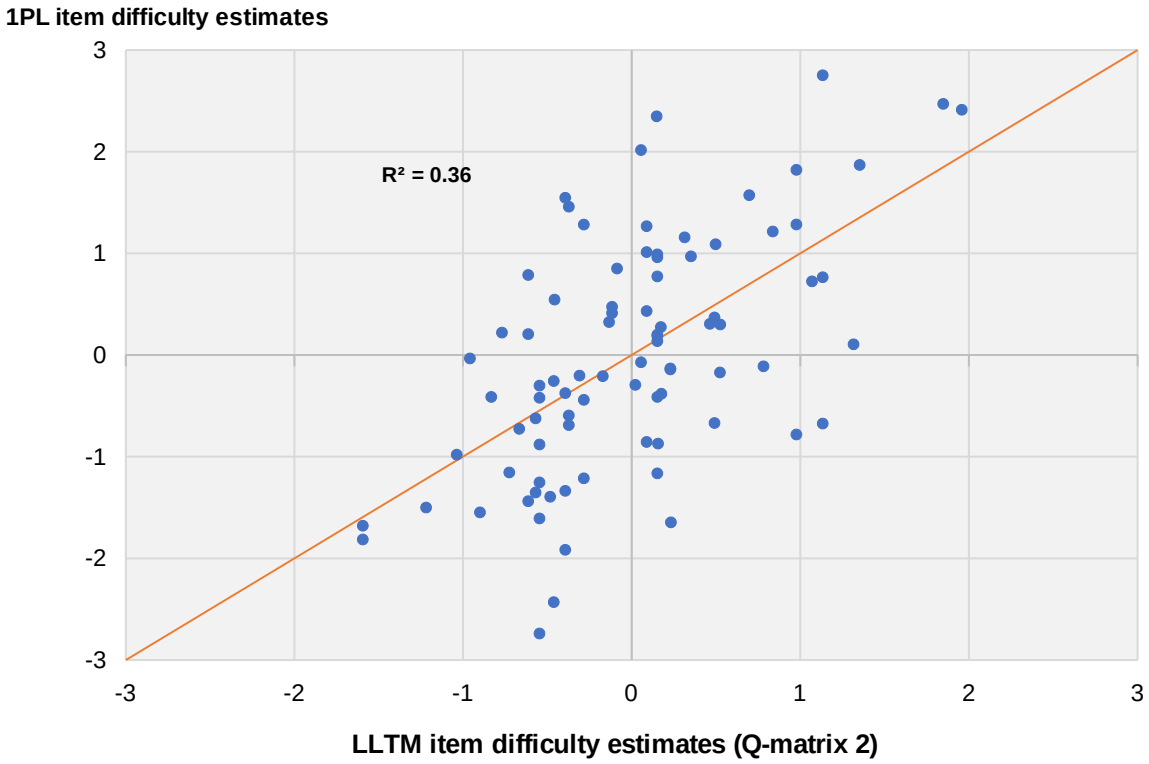
The correlation between 1PL and LLTM estimates for Q-matrix 2 was moderate ($r = .60, r^2 = .36$). This relationship is presented graphically in Figure 19. By comparison, Q-matrix D explained a greater proportion of the variance in 1PL item difficulty estimates. This indicates that there *was* some loss of model fit caused by the reduction of attributes from Q-matrix D to Q-matrix 2.

Table 33: Eta difficulty estimates for attributes in Q-matrix 2

Item attribute	Eta parameter	*SE*	95% Lower CI	95% Upper CI	*p*
j_{201} – Whole numbers, fractions and decimals	0	-	-	-	-
j_{202} – Geometry	-0.063	0.010	-0.083	-0.043	<.001***
j_{203} – Data	-1.196	0.017	-1.228	-1.163	<.001***
j_{204} – Measuring or estimating in real-world contexts	0.339	0.014	0.312	0.367	<.001***
j_{205} – Perform computations in arithmetic / geometry	0.635	0.009	0.617	0.654	<.001***
j_{206} – Make judgements or evaluate correctness	0.375	0.010	0.355	0.396	<.001***
j_{207} – Logical reasoning in verbal or visual contexts	0.981	0.009	0.962	1.000	<.001***
j_{208} – Number sense / place-value awareness	0.199	0.012	0.175	0.224	<.001***
j_{209} – Patterns and sequences	-0.268	0.014	-0.296	-0.240	<.001***
j_{210} – Proportional relationships	1.460	0.020	1.420	1.500	<.001***
j_{211} – Compare entities	0.085	0.011	0.063	0.107	<.001***
j_{212} – Checking/evaluating options	-0.286	0.011	-0.307	-0.265	<.001***

Eta parameters not calculated for one attribute due to Q-matrix structure.
CI = confidence interval. * = sig. at 95% confidence level, ** = sig at 99% confidence, *** = sig. at 99.9% confidence

Figure 19: Scatterplot of 1PL and LLTM item difficulty estimates (Q-matrix 2)



Appendix F – Guessing and slipping parameters

Table 34 presents the item-level guessing and slipping parameters of each item under Q-matrix 1 and Q-matrix 2 in paper 3.

Table 34: Item-level guessing and slipping parameters

Item ID	Q-matrix 1		Q-matrix 2	
	Guessing (g_i)	Slipping (s_i)	Guessing (g_i)	Slipping (s_i)
M031346A	0.4258	0.017	0.446	0.018
M031346B	0.0967	0.2788	0.060	0.321
M031346C	0.0265	0.4495	0.002	0.460
M031379	0.105	0.4903	0.095	0.534
M031380	0.058	0.6155	0.053	0.605
M031313	0.5253	0.1382	0.514	0.116
M031083	0.6939	0.0135	0.723	0.000
M031071	0.251	0.2812	0.270	0.243
M031185	0.281	0.1976	0.194	0.083
M051305	0.4092	0.1269	0.416	0.069
M051091	0.3098	0.1297	0.328	0.079
M051001	0.0936	0.3195	0.054	0.360
M051007	0.2451	0.4943	0.230	0.486
M051203	0.0364	0.793	0.026	0.781
M051601	0.3935	0.1691	0.392	0.077
M051064A	0.4609	0.1109	0.415	0.104
M051064B	0.9444	0.0097	0.939	0.008
M051015	0.2527	0.235	0.304	0.197
M051123	0.346	0.2951	0.315	0.286
M051109	0.2028	0.0583	0.314	0.036
M051117	0.3634	0.2083	0.447	0.165
M041010	0.6081	0.0825	0.596	0.063
M041098	0.3817	0.2588	0.365	0.253
M041064	0.5556	0.0083	0.447	0.041
M041003	0.3291	0.2272	0.319	0.157
M041104	0.3758	0.061	0.301	0.058
M041299	0.0585	0.2763	0.056	0.274
M041329	0.35	0.1552	0.321	0.172
M041143	0.4797	0.164	0.403	0.210
M041158	0.5672	0.0932	0.602	0.076
M041328	0.4241	0.0918	0.288	0.113
M041155	0.2648	0.2247	0.263	0.190
M041284	0.0089	0.73	0.012	0.763
M041335	0.4468	0.0986	0.512	0.077
M041184	0.654	0.0454	0.725	0.040
M051205	0.1604	0.3789	0.159	0.395
M051039	0.2895	0.1489	0.284	0.175
M051055	0.0305	0.7278	0.014	0.700
M051006	0.0264	0.7616	0.025	0.784
M051070	0.1801	0.5911	0.179	0.586

M051018	0.3467	0.3254	0.347	0.335
M051407	0.4415	0.2212	0.469	0.175
M051410	0.1731	0.3424	0.173	0.328
M051059	0.6781	0.0155	0.586	0.022
M051093	0.2748	0.199	0.312	0.306
M051134	0.0898	0.2437	0.109	0.216
M051077	0.2462	0.2025	0.278	0.199
M031128	0.3413	0.2447	0.344	0.262
M031016	0.0787	0.3928	0.046	0.422
M031183	0.0677	0.3067	0.057	0.318
M031187	0.681	0.0511	0.682	0.046
M031251	0.2634	0.3196	0.264	0.300
M031294	0.2341	0.0871	0.317	0.063
M031297	0.174	0.3205	0.191	0.302
M031218	0.2952	0.1972	0.245	0.198
M031109	0.4671	0.1303	0.489	0.132
M031159	0.6021	0.0488	0.621	0.034
M031133	0.7932	0.0464	0.772	0.042
M041107	0.7266	0.0333	0.718	0.024
M041011	0.5256	0.0696	0.500	0.060
M041122	0.057	0.5113	0.044	0.503
M041041	0.4862	0.201	0.483	0.182
M041320	0.3026	0.105	0.292	0.100
M041115A	0.5	0.1624	0.499	0.127
M041115B	0.2371	0.1435	0.219	0.153
M041160A	0.7897	0.0197	0.719	0.030
M041160B	0.9432	0.0027	0.919	0.002
M041327	0.5782	0.1001	0.481	0.112
M041148	0.195	0.319	0.200	0.332
M041265	0.2191	0.3483	0.253	0.450
M041175	0.7446	0.0184	0.742	0.019
M041199	0.5563	0.0105	0.638	0.022
M031210	0.3063	0.236	0.339	0.202
M031009	0.2853	0.2022	0.266	0.167
M031252	0.5595	0.0755	0.553	0.052
M031316	0.5969	0.0612	0.524	0.054
M031317	0.1439	0.5462	0.119	0.567
M031079B	0.5844	0.0518	0.568	0.033
M031079C	0.2504	0.4264	0.228	0.422
M031004	0.0837	0.6201	0.100	0.562
M031043	0.2563	0.2564	0.259	0.214
M031325	0.1629	0.355	0.149	0.341
M031088	0.6634	0.089	0.665	0.089
M031093	0.6943	0.1275	0.718	0.144
M031155	0.2673	0.1495	0.321	0.165