Flinders UNIVERSITY
ADELAIDE · SOUTH AUSTRALIA

Local item independence in large scale international assessments: the Programme for International Student Assessment (PISA) case

by

Pawel Piotr Skuza

Thesis
Submitted to Flinders University
for the degree of

Doctor of Philosophy

College of Education, Psychology and Social Work

June 2018

Principal Supervisor: **Associate Professor David Curtis**

Associate Supervisor: **Professor Rosalind Murray-Harvey**

Adjunct Supervisor: **Professor Kelvin Gregory**

Table of Contents

ABSTRACT

International large-scale comparative educational assessments have a 50-year history, and currently, two major international, multidisciplinary, longitudinal, large-scale assessments are implemented: the Trends in International Mathematics and Science Study (TIMSS) and the Programme for International Student Assessment (PISA). PISA, the study of interest in this thesis, focuses on students' ability to apply their knowledge and it makes greater use of testlets, i.e. a group of items with a common stimulus, than does TIMSS. PISA uses a generalised form of the Rasch model for scaling cognitive data, and this measurement model makes various underlying assumptions. One of these is "local item independence" (LII), meaning that after controlling for students' ability, items in the assessment should be independent of each other. In practice, independence should be revealed by low residual correlations between items after modelling student ability. All measures are subject to error, and a challenge for educational testing is to minimise that error. Violating the assumption of item independence may increase measurement error, and in this thesis, violations of item independence in PISA are examined.

Three main research aims are investigated in this thesis. The first aim is to describe the testlets used in PISA by providing an overview of items and testlets used across multiple PISA waves. The second research aim is to examine data from PISA's international calibration samples for the existence of local item dependence (LID) by utilising a non-IRT based LID index namely "Residual Correlation from Factor Analysis". Meta-analysis is used to combine estimates of LID prevalence across multiple PISA waves, and multilevel logistic regression is used to investigate which item characteristics predict the presence of item dependency. An in-depth investigation of the LID drivers for released items is offered. The cross-wave consistency in LID presence is reported. The third aim is to examine LID within the national calibration level data aiming to identify countries with a higher level of local item dependency. The cross-national consistency of LID existence is investigated along with consideration of the possibility of differential testlet functioning. Greater incidence of LID in some national samples may reveal country-specific causes of LID and they could arise from differences in curriculum and pedagogies or in the administration of the tests.

Results reveal that the reading assessment in PISA makes greater use of multi-item testlets than occurs in the mathematics or science assessments. Single-item testlets are more common in the mathematics assessments than in reading or science. In the investigation of LID, both positive and negative residual correlations were found. Positive residual correlations are expected for items that, for example, share a common prompt, but negative residual correlations are also found and possible

causes for both are suggested. Analyses of international calibration data reveal that positive item dependence is as prevalent in mathematics as it is in reading, with science showing less item dependence. Although within-testlet positive LID is present, pairs of items from different testlets also show positive LID and utilising publicly released items allow this between-testlet dependency to be examined and explained. Some testlets exhibit positive LID among the majority of their items, yet other testlets do not indicate within-testlet LID despite having a shared stimulus. Item dependency is shown to be consistently present for some testlets across all PISA waves in which they were used for the purpose of cross-wave linking. Negative dependence is more prevalent than positive LID. Plausible drivers of positive LID are offered for item pairs which come from released items. While it is often assumed that the use of a common item prompt is responsible for LID, multilevel logistic models point to other drivers of positive LID. For example in mathematics, the difference in items' difficulty or item pair mathematical strand are associated with positive LID. The specific skill of being able to offer a scientific judgement drives some of the dependency apparent in science literacy. While the study offered some signs that selective time and effort allocation could drive a negative LID as suggested by Yen (1993), negative dependency was more likely a mathematical artefact of positive within-testlet LID as proposed by Habing and Roussos (2003). Differences in the prevalence of LID among the 24 investigated OECD countries indicate that Greece more frequently showed high levels of positive within-testlet dependency while between-testlet positive LID was greater for high performing countries such as Finland, Japan and Korea. LID investigations when international and national calibration datasets are used reveals a consistency in dependency between countries for some testlets but also suggests the possibility of differential testlet functioning for others.

Findings from this research are applicable to all educational assessments that use testlets as a part of their cognitive skills testing and for PISA test development teams in particular. Closer consideration should be given to possible within-testlet as well as a between-testlet dependency at the stage of preparing and field testing cognitive items. The need for more research on the effects of LID on cross-wave linking is warranted. Practical implications and suggestions for future research are given.

In conclusion, this research provided evidence that local item independence is violated in PISA and a range of plausible causes are identified. The research has extended the limited literature about LID in PISA to provide a broader perspective utilising data from three cognitive domains and five waves of PISA. The generalisability of findings were strengthened by showing cross-wave and cross national consistency in LID presence.

DECLARATION

I certify that this thesis does not incorporate without acknowledgment any material previously submitted for a degree or diploma in any university, and that to the best of my knowledge and belief it does not contain any material previously published or written by another person except where due reference is made in the text.

ACKNOWLEDGEMENTS

I am grateful to have Associate Professor David Curtis as my principal supervisor whose knowledge, patience and encouragement were very important in the last few years of my part-time doctoral candidature. He not only helped to see this research to be completed but also taught me many traits needed to be an excellent supervisor. I am also grateful for the ongoing support offered by my associate supervisor Professor Rosalind Murray-Harvey and adjunct supervisor Professor Kelvin Gregory.

I am thankful for the patience of my wife during my part-time studies. Thank you my love. Big hugs go to my three children for being my sunshine and to my father for words of encouragement.

My balance of full-time work and study was supported by my work managers (Mark Legg, Dean Gawler, Amanda Nixon and Liz Walkley Hall). I was also aided in my research by Sean Reilly from eResearch SA High-Performance Computing team who offered clarifications on using HPC. Sophie Vayssettes from OECD along with dr Michael Timms and Alla Berezner from Australian Council for Educational Research kindly addressed questions I had about the PISA methodology. Finally, Jeni Thomas contributed to English proofreading of the final draft of the thesis according to the conditions laid out in the university endorsed national guidelines.

This research was supported by an Australian Government Research Training Program (RTP) Scholarship and its first year also by an Australian Postgraduate Award scholarship.

LIST OF TABLES

LIST OF FIGURES

LIST OF ELECTRONIC APPENDICES

- [1]Electronic Appendix 3.3.1 - IBM SPSS syntax code to extract various statistics from the Mplus CFA outputs using PISA's data as example.sps
- Electronic Figure 4.3.1.pdf
- Electronic Figure 4.3.1_Alphabetic order.pdf
- Electronic Figure 4.3.2.pdf
- Electronic Figure 4.3.2_Alphabetic order.pdf
- Electronic Figure 4.3.3.pdf
- Electronic Figure 4.3.3_Alphabetic order.pdf
- Electronic Appendix for Table 5.4.1 - Mathematics - Positive LID.xlsx
- Electronic Appendix for Table 5.4.2 - Reading - Positive LID.xlsx
- Electronic Appendix for Table 5.4.3 - Science - Positive LID.xlsx
- Electronic Appendix for Table 5.4.4 - Mathematics - Negative LID.xlsx
- Electronic Appendix for Table 5.4.5 - Reading - Negative LID.xlsx
- Electronic Appendix for Table 5.4.6 - Science - Negative LID.xlsx
- Electronic Figure 5.4.2 - Mathematics PISA 2000.pdf
- Electronic Figure 5.4.2 - Mathematics PISA 2000_ALL RCs.pdf
- Electronic Appendix for Figure 5.4.2_POSTIVE LID_WITHIN TESTLET - PISA 2000 Mathematics.xlsx
- Electronic Appendix for Figure 5.4.2_POSTIVE LID_BETWEEN TESTLETS - PISA 2000 Mathematics.xlsx
- Electronic Appendix for Figure 5.4.2_NEGATIVE LID_BETWEEN TESTETS - PISA 2000 Mathematics.xlsx
- Electronic Figure 5.4.3 - Mathematics PISA 2003.pdf
- Electronic Figure 5.4.3 - Mathematics PISA 2003_ALL RCs.pdf
- Electronic Appendix for Figure 5.4.3_POSTIVE LID_WITHIN TESTLET - PISA 2003 Mathematics.xlsx
- Electronic Appendix for Figure 5.4.3_POSTIVE LID_BETWEEN TESTLETS - PISA 2003 Mathematics.xlsx
- Electronic Appendix for Figure 5.4.3_NEGATIVE LID_BETWEEN TESTLETS - PISA 2003 Mathematics.xlsx
- Electronic Figure 5.4.4 - Mathematics PISA 2006.pdf
- Electronic Figure 5.4.4 - Mathematics PISA 2006_ALL RCs.pdf
- Electronic Appendix for Figure 5.4.4_POSTIVE LID_WITHIN TESTLET - PISA 2006 Mathematics.xlsx
- Electronic Appendix for Figure 5.4.4_POSTIVE LID_BETWEEN TESTLETS - PISA 2006 Mathematics.xlsx
- Electronic Appendix for Figure 5.4.4_NEGATIVE LID_BETWEEN TESTLETS - PISA 2006 Mathematics.xlsx
- Electronic Figure 5.4.5 - Mathematics PISA 2009.pdf
- Electronic Figure 5.4.5 - Mathematics PISA 2009_ALL RCs.pdf
- Electronic Appendix for Figure 5.4.5_POSTIVE LID_WITHIN TESTLET - PISA 2009 Mathematics.xlsx
- Electronic Appendix for Figure 5.4.5_POSTIVE LID_BETWEEN TESTLETS - PISA 2009 Mathematics.xlsx

[1] In order for the hyperlinks to the files to work correctly, the folder called *Electronic Appendices* containing all the files needs to be placed in the same folder that the thesis file is located.

- Electronic Appendix for Figure 5.4.5_NEGATIVE LID_BETWEEN TESTLETS - PISA 2009 Mathematics.xlsx
- Electronic Figure 5.4.6 - Mathematics PISA 2012.pdf
- Electronic Figure 5.4.6 - Mathematics PISA 2012_ALL RCs.pdf
- Electronic Appendix for Figure 5.4.6_POSTIVE LID_WITHIN TESTLET - PISA 2012 Mathematics.xlsx
- Electronic Appendix for Figure 5.4.6_POSTIVE LID_BETWEEN TESTLETS - PISA 2012 Mathematics.xlsx
- Electronic Appendix for Figure 5.4.6_NEGATIVE LID_WITHIN TESTLET - PISA 2012 Mathematics.xlsx
- Electronic Appendix for Figure 5.4.6_NEGATIVE LID_BETWEEN TESTLETS - PISA 2012 Mathematics.xlsx
- Electronic Appendix for Figures 5.4.2-6 - MATHEMATICS.xlsx
- Electronic Figure 5.4.7 - Reading PISA 2000.pdf
- Electronic Figure 5.4.7 - Reading PISA 2000_ALL RCs.pdf
- Electronic Appendix for Figure 5.4.7_POSTIVE LID_WITHIN TESTLET - PISA 2000 Reading.xlsx
- Electronic Appendix for Figure 5.4.7_POSTIVE LID_BETWEEN TESTLETS - PISA 2000 Reading.xlsx
- Electronic Appendix for Figure 5.4.7_NEGATIVE LID_BETWEEN TESTLETS - PISA 2000 Reading.xlsx
- Electronic Figure 5.4.8 - Reading PISA 2003.pdf
- Electronic Figure 5.4.8 - Reading PISA 2003_ALL RCs.pdf
- Electronic Appendix for Figure 5.4.8_POSTIVE LID_WITHIN TESTLET - PISA 2003 Reading.xlsx
- Electronic Appendix for Figure 5.4.8_NEGATIVE LID_BETWEEN TESTLETS - PISA 2003 Reading.xlsx
- Electronic Figure 5.4.9 - Reading PISA 2006.pdf
- Electronic Figure 5.4.9 - Reading PISA 2006_ALL RCs.pdf
- Electronic Appendix for Figure 5.4.9_POSTIVE LID_WITHIN TESTLET - PISA 2006 Reading.xlsx
- Electronic Appendix for Figure 5.4.9_NEGATIVE LID_BETWEEN TESTLETS - PISA 2006 Reading.xlsx
- Electronic Figure 5.4.10 - Reading PISA 2009.pdf
- Electronic Figure 5.4.10 - Reading PISA 2009_ALL RCs.pdf
- Electronic Appendix for Figure 5.4.10_POSTIVE LID_WITHIN TESTLET - PISA 2009 Reading.xlsx
- Electronic Appendix for Figure 5.4.10_POSTIVE LID_BETWEEN TESTLETS - PISA 2009 Reading.xlsx
- Electronic Appendix for Figure 5.4.10_NEGATIVE LID_BETWEEN TESTLETS - PISA 2009 Reading.xlsx
- Electronic Figure 5.4.11 - Reading PISA 2012.pdf
- Electronic Figure 5.4.11 - Reading PISA 2012_ALL RCs.pdf
- Electronic Appendix for Figure 5.4.11_POSTIVE LID_WITHIN TESTLET - PISA 2012 Reading.xlsx
- Electronic Appendix for Figure 5.4.11_POSTIVE LID_BETWEEN TESTLETS - PISA 2012 Reading.xlsx
- Electronic Appendix for Figure 5.4.11_NEGATIVE LID_BETWEEN TESTLETS - PISA 2012 Reading.xlsx
- Electronic Appendix for Figures 5.4.7-11 - READING.xlsx

- Electronic Figure 5.4.12 - Science PISA 2000.pdf
- Electronic Figure 5.4.12 - Science PISA 2000_ALL RCs.pdf
- Electronic Appendix for Figure 5.4.12_POSTIVE LID_WITHIN TESTLET - PISA 2000 Science.xlsx
- Electronic Appendix for Figure 5.4.12_POSTIVE LID_BETWEEN TESTLETS - PISA 2000 Science.xlsx
- Electronic Appendix for Figure 5.4.12_NEGATIVE LID_BETWEEN TESTLETS - PISA 2000 Science.xlsx
- Electronic Figure 5.4.13 - Science PISA 2003.pdf
- Electronic Figure 5.4.13 - Science PISA 2003_ALL RCs.pdf
- Electronic Appendix for Figure 5.4.13_POSTIVE LID_WITHIN TESTLET - PISA 2003 Science.xlsx
- Electronic Appendix for Figure 5.4.13_POSTIVE LID_BETWEEN TESTLETS - PISA 2003 Science.xlsx
- Electronic Appendix for Figure 5.4.13_NEGATIVE LID_BETWEEN TESTLETS - PISA 2003 Science.xlsx
- Electronic Figure 5.4.14 - Science PISA 2006.pdf
- Electronic Figure 5.4.14 - Science PISA 2006_ALL RCs.pdf
- Electronic Appendix for Figure 5.4.14_POSTIVE LID_WITHIN TESTLET - PISA 2006 Science.xlsx
- Electronic Appendix for Figure 5.4.14_POSTIVE LID_BETWEEN TESTLETS - PISA 2006 Science.xlsx
- Electronic Appendix for Figure 5.4.14_NEGATIVE LID_BETWEEN TESTLETS - PISA 2006 Science.xlsx
- Electronic Figure 5.4.15 - Science PISA 2009.pdf
- Electronic Figure 5.4.15 - Science PISA 2009_ALL RCs.pdf
- Electronic Appendix for Figure 5.4.15_POSTIVE LID_WITHIN TESTLET - PISA 2009 Science.xlsx
- Electronic Appendix for Figure 5.4.15_POSTIVE LID_BETWEEN TESTLETS - PISA 2009 Science.xlsx
- Electronic Appendix for Figure 5.4.15_NEGATIVE LID_BETWEEN TESTLETS - PISA 2009 Science.xlsx
- Electronic Figure 5.4.16 - Science PISA 2012.pdf
- Electronic Figure 5.4.16 - Science PISA 2012_ALL RCs.pdf
- Electronic Appendix for Figure 5.4.16_POSTIVE LID_WITHIN TESTLET - PISA 2012 Science.xlsx
- Electronic Appendix for Figure 5.4.16_POSTIVE LID_BETWEEN TESTLETS - PISA 2012 Science.xlsx
- Electronic Appendix for Figure 5.4.16_NEGATIVE LID_BETWEEN TESTLETS - PISA 2012 Science.xlsx
- Electronic Appendix for Figures 5.4.12-16 - SCIENCE.xlsx
- Electronic Appendix 7.2.1_List of all CFA warnings produced from national calibration data.xlsx
- Electronic Appendix 7.2.2_List of all CFA warnings produced from international calibration data.xlsx
- Electronic Appendix 7.2.3_Selected cross tabulations for mathematical data from France in PISA 2000.xlsx

LIST OF ACRONYMS

ACE - American Council on Education

AERA - American Educational Research Association

AIC - Akaike Information Criterion

ANOVA - Analysis of variance

APA - American Psychological Association

BIB - Balanced Incomplete Block

BIC - Bayesian information criterion

CAT - Computer Adaptive Testing

CCFA - Categorical Confirmatory Factor Analysis

CFA - Confirmatory Factor Analysis

CFI - Comparative Fit Index

CI - Confidence Interval

CMC - Complex Multiple Choice

CTT - Classical Test Theory

DIF - Differential Item Functioning

ESD - Extreme Studentize Deviate

FA - Factor Analysis

HPC - High-Performance Computing

ICC - Intraclass Correlation Coefficient

IRT - Item Response Theory

LID - Local Item Dependence

LII - Local Item Independence

MAT - Multidimensional Adaptive Testing

MCML - Mixed-Coefficients Multinomial Logit model

MI - Modification Indices

MMS - Multiple matrix sampling

NCES - National Center for Education Statistics

NCME - National Council on Measurement in Education

NEAP - National Assessment of Educational Progress

OCR - Open Constructed Response

OECD - Organisation for Economic Co-operation and Development

OR - Odds Ratio

PCA - Principal Component Analysis

PISA - The Programme for International Student Assessment

RC - Residual Correlation

RMSEA - Root Mean Square Error of Approximation

SEM - Standard Error of Measurement

SEM - Structural Equation Modelling

SLD - Surface Local Item Dependence

SMC - Simple Multiple Choice

SR - Short Response

TIMSS - Trends in International Mathematics and Science Study

TLI - Tucker Lewis Index

ULD - Underlying Local Item Dependence

WLSMV - Weighted Least Squares Mean and Variance adjusted estimator

WRMR - Weighted Root-Mean-square Residual

CHAPTER 1 INTRODUCTION

The research for this thesis is an investigation of the presence and plausible causes of local item dependence (LID) in the Programme for International Student Assessment (PISA) study. Problems with the quality of the measurement of mathematical, scientific and reading achievement might arise when the assumption - that each item is independent of other items after controlling for underlying latent cognitive ability - is violated. This is the assumption of Local Item Independence (LII). Where this assumption is violated, Local Item Dependency (LID) is observed. Constructing efficient tests gives preference towards the use of testlets of items based on a common prompt. This chapter sets up the context of the study; describes the purpose and significance of the research; outlines the research aims and corresponding objectives; and elaborates how this study adds to the body of the research about LID presence in PISA.

1.1 Context of the study

1.1.1 Large-scale educational assessments and their impact on educational systems

For over 50 years, the majority of developed countries have been participating in international large-scale comparative assessments (Johnson, 1999; Kamens & McNeely, 2010; Kellaghan & Greaney, 2001). Furthermore, the number of developing countries taking part in worldwide testing programmes has also been increasing (Bloem, 2015; Milford, Ross, & Anderson, 2010; Wagner, 2010). The drive towards assessing skills and learning outcomes has been to some extent orchestrated by various international organisations such as the World Bank and the OECD (Akkari & Lauwerier, 2015; McGaw, 2008). Numerous international large-scale assessments have been conducted. Grisay and Griffin (2006) reported more than 20 major cross-national assessments and provided comprehensive information about them. However, there are two major multidisciplinary, longitudinal, international large-scale assessments currently implemented: the Trends in International Mathematics and Science Study (TIMSS) and the Programme for International Student Assessment (PISA). Succinct overviews of TIMSS and PISA have been prepared by Robitaille and Beaton (2002) and Turner and Adams (2007), respectively, with a cross comparison overview offered by Kell and Kell (2014a). An increasing number of countries have been taking part in both studies in

the subsequent data collections for both longitudinal testing cycles. For example, in the PISA 2012 survey, data were collected from just over half a million students, representing about 28 million 15 year-olds in the schools of the 65 participating countries and economies (OECD, 2014b). This figure is close to twice as many participants and countries as compared with the initial group assessed in the year 2000 (Adams & Wu, 2002).

The impact of large-scale comparative assessments has been widely recognised to include their influence on educational research, policy making, and on teaching and curriculum (Best et al., 2013; Care, Griffin, Zhang, & Hutchinson, 2014; Elley, 2005; Gilmore, 2005; Hopfenbeck et al., 2017; Naumann, 2005; Pons, 2017; von Davier, 2013; Wagemaker, 2004; Wiseman, 2010). In particular, the effect of the PISA study has been investigated in a cross-national comparison framework (Baird et al., 2011; Breakspear, 2012; Martens & Niemann, 2013; Meyer & Benavot, 2013; Yore, Anderson, & Chiu, 2010). PISA's influence is also addressed in various publications targeting the specific national impact that the PISA study has had (Arzarello, Garuti, & Ricci, 2015; Białecki, Jakubowski, & Wiśniewski, 2017; Carvalho, Costa, & Gonçalves, 2017; Ferrer, 2017; Gorur & Wu, 2015; Tasaki, 2017; Volante, 2013; Yang & Lin, 2015). Its impact on educational research has also been investigated (Anderson, Lin, Treagust, Ross, & Yore, 2007; Domínguez, Vieira, & Vidal, 2012; Prenzel, Kobarg, Schöps, & Rönnebeck, 2013). Finally, each release of PISA results brings about considerable international (including Australia-wide) media coverage (Fladmoe, 2012; Hopfenbeck & Görgen, 2017; Waldow, Takayama, & Sung, 2014). Consequently, studies such as PISA and TIMSS can be regarded, to different degrees, as high-stakes assessments for the bodies that govern educational systems in countries participating in those international evaluations. Intra-national and international educational performance comparisons are made based on the large-scale educational assessments summary data and policy actions are taken based on this data (Lietz & Tobin, 2016; Michel, 2017).

1.1.2 Importance of assumptions underlying mathematical models used in creating the educational proficiency scales

The need for highly reliable and valid tests has been clearly highlighted over time in successive editions of the *Standards for educational and psychological testing* (AERA, APA, & NCME, 1974, 1985, 1999) and *Educational Measurement* (Brennan, 2006b; Linn, 1989). As mentioned above, and elaborated further in the relevant sections of the Literature Review, various item response theory (IRT) models exist, and some of them are employed in

international large-scale educational assessments. Despite the considerable theoretical progress that has been made in measuring students' knowledge and the availability of computers that have facilitated implementation of very sophisticated measurement models, many prominent psychometricians call for careful consideration of the limitations of these models. Brennan's (2007) influential text outlines the types of issues that researchers are concerned with:

> Given the inconsistencies across models, it is natural to ask which model provides the correct or right answer to the questions posed. For the most part, there is no right answer, and investigators searching for that "Holy Grail" will be forever disappointed. The models are just that – models, not reality; each of them has its own set of definitions and assumptions, and the definitions and assumptions do not mesh perfectly across models. (Brennan, 2006b, p. 7)

Given the complexity of PISA, there is a need to address various threats to the robustness of its results. The PISA 2012 version of the manual (OECD, 2014b) approaches 500 pages. In its various chapters, it deals with many IRT related design challenges, including sampling students, translation of survey materials, implementing procedures for monitoring quality, deriving weights and undertaking data management and reliability studies. Each PISA technical manual also includes a chapter dedicated to scaling its cognitive data. All five waves of PISA investigated in this research used a generalised form of the Rasch model, called mixed coefficients multinomial logit model (Adams, Wilson, & Wang, 1997; Adams & Wu, 2007). This IRT model is used in each PISA study in three steps: national calibrations, international scaling and student score generation. As with any IRT model, it is essential that the assumptions, sometimes referred by some psychometricians as requirements, are evaluated and followed (Yen & Fitzpatrick, 2006).

1.1.3 Brief overview of IRT assumptions

There are three main assumptions of unidimensional IRT models, namely monotonicity, unidimensionality and local item independence (LII) (De Ayala, 2009, p. 20; Nandakumar & Ackerman, 2004, p. 93). While relevant sections in the Literature Review elaborate on all three assumptions further, a simplified definition of LII is briefly introduced here. The local item independence assumption requires that after controlling for the examinee's inferred ability on the measure of interest, item responses should be independent (Yen & Fitzpatrick, 2006, p. 122). Although there may be various situations when this assumption is violated, one of the most commonly mentioned in the literature examples occurs when students respond to multiple questions that are based on a common stimulus (i.e. reading passage dependence).

Violation of LII may take place in this case because those questions stemmed from the same reading passage, and to particular prior knowledge related to the common passage may make items dependent on each other. The existence of LID undermines the quality of the measures, and its specific consequences are detailed in section 2.6.

1.2 Overall purpose of the study

So far it has been argued that PISA is a high-stakes policy-relevant study, yet a considerable number of technical limitations are alluded to by various authors. In light of the highlighted importance of maintaining sufficient compliance to requirements of the mathematical measurement models, the overall purpose of this study is to investigate the adherence of PISA's cognitive data to the LII assumption. This purpose is of scientific relevance as the research investigating the LII assumption in PISA is scarce. At the time when this part-time research project was conceived there appeared not to be a single scientific reference relating to the assessment of whether local item independence (LII) was violated in the PISA study. During the last six years, a few publications have appeared that address this issue (Kreiner, 2011; Lyons-Thomas, Sandilands, & Ercikan, 2014; Monseur, Baye, Lafontaine, & Quittre, 2011; Oliden & Lizaso, 2014; Trendtel, Ünlü, Kasper, & Stubben, 2014). However, in all research papers and reports dealing with LII in PISA, it is tested only for a specific wave, domain and/or country without a more generalised cross-wave or cross-country description being being pursued. This research expands the scope of listed above publications by including data from five PISA studies, three cognitive domains and twenty-four national datasets. None of the PISA technical manuals (Adams & Wu, 2002; OECD, 2005b, 2009b, 2012, 2014b) acknowledged that LII was evaluated in either national or international data calibration stages. Furthermore, the need for attention to be given to local item dependence (LID) was one of the recommendations put forward by Mazzeo and von Davier (2008) in an OECD-sponsored audit of the PISA test design.

In this thesis, evidence of violations of LII in PISA is sought, and where it is found, the patterns of LID prevalence and plausible causes are investigated.

1.3 Research aims and objectives

There are three main aims which underly this research project.

Firstly, an intention is to provide a comprehensive overview of the testlets used in PISA.

This aim will be addressed by looking at each of three cognitive domains (i.e. mathematics, reading and science) in regards to its longitudinal patterns of testlets use across five PISA studies. Particular attention will be given to within-testlet variability of items' difficulties.

Secondly, the study aims to investigate the existence of LID in data from PISA's international calibrations. This aim will be fulfilled by reporting the positive and negative LID prevalence by utilising a non-IRT based LID index. Differences in the prevalence of LID across the cognitive domain, PISA wave, and cross wave linking will be investigated. Plausible reasons for any LID will be investigated by taking advantage of released items and other information about the cognitive items used in PISA. The consistency of LID presence in testlets used for cross-wave PISA linking will be looked at.

Thirdly, as a final aim of the research, LID in national datasets will be evaluated. This aim will be addressed by comparing national test data against the international level results. The difference in LID presence among the countries will be highlighted aiming also to identify differential testlet functioning.

Detailed research questions corresponding to the objectives listed above, are introduced in the methodology chapter (see Section 3.1).

1.4 Significance of the study

The scientific understanding of the prevalence and characteristics of LID in the PISA study is limited as reviewed in section 2.9. However, an evaluation of the LID assumption has been given considerable attention in many applied papers from various research fields such as education (Natesan & Kieftenbeld, 2013), psychology (de Klerk, Nel, Hill, & Koekemoer, 2013; Lundgren-Nilsson, Jonsdottir, Ahlborg, & Tennant, 2013), medicine (Anatchkova et al., 2014; Hamilton et al., 2015; Kisala et al., 2015), disability (Kent, Grotle, Dunn, Albert, & Lauridsen, 2015; Wang, Hart, Deutscher, Yen, & Mioduski, 2013) and nursing (Kaspar & Hartig, 2016). This thesis seeks to fill a gap in our knowledge about the psychometric properties of the PISA data related to adherence to the local item independence assumption.

Should the existence of LID in PISA be confirmed by this research, this may suggest the need to modify the test development procedures. The results could also contribute to the discussion on whether testlet-based IRT models should be used in future implementations of

this large-scale educational assessment.

Each PISA technical manual gives some information about individual items. However, information about the allocation of items to testlets has to be extracted manually from the headings of the item names. Arguably no publication looks at the longitudinal patterns of cognitive questions and testlet utilisation that is covered in this thesis.

Conditional upon the magnitude and pattern of LID detected it is anticipated that this research may initiate subsequent research investigations into, for example, the impact of LID on the size of the standard errors in the cross-national comparisons and therefore into the accuracy of parameter estimates and possibly of countries' estimated means.

1.5 Organization of the thesis

This thesis is organised as follows.

Chapter one sets up the context and purpose of the thesis. An overview of what is known and not known about LID in PISA is presented, and the significance of the proposed research is explained. Three aims and corresponding research objectives are outlined.

Chapter two reviews the relevant literature. It starts with a succinct overview of educational measurement, and it offers an introduction of the IRT models and their assumptions leading to a summary of students' abilities' scaling as it is implemented in the PISA study. PISA's merits are reviewed next. This is followed by a discussion of the definition of LID through to an overview of possible reasons behind LID presence. This part of the chapter is followed by sections addressing the consequences of LID presence and strategies for managing it. The second half of the chapter reviews various methods used for LID detection. The chapter concludes with a review of reported LID occurrences with a particular focus on large-scale international assessments in general and PISA in particular.

Chapter three describes the research design and it begins by listing detailed research questions. As each of the three research aims utilised datasets collated purposively to address each aim the chapter is separated into corresponding sections. The organisation of each subchapter starts with the description of data and software used followed by a plan for data analysis. The sections also discuss assumptions of statistical approaches utilised.

Chapters four, five and six present the results and interpretation of the analyses that are

undertaken. Once again the organisation of these chapters adheres to the order of general research aims followed by specific research questions.

Chapter seven discusses the key findings and addresses limitations. This chapter also elaborates on possible practical consequences for conducting future PISA studies, and possibilities for future research.

CHAPTER 2 REVIEW OF THE LITERATURE

This chapter starts with a brief overview of educational measurement and a historical synopsis of test theory moving into an introduction to the methodology used in scaling cognitive data in a PISA study. These two initial sections serve the purpose of outlining the mathematical models which require an assumption of Local Item Independence (LII). After an introduction to the PISA study, a section summarising a debate about the merits of PISA is presented. The definition of local item independence is then offered, followed by a review of different taxonomies of violations of LII, i.e. local item dependency (LID). Plausible causes of LID are identified which would be investigated in the PISA study by this research. The consequences of local item dependence are reviewed with a short section discussing methods used to manage the presence of local dependence. Methods of detecting LID are reviewed with a particular focus on the index chosen for this study. Documented occurrences of LID in large scale educational assessments are reported with a particular focus on publications about the PISA study. The chapter concludes with a short summary suggesting which gaps in the knowledge about LID in PISA this research addresses.

2.1 Brief introduction to educational measurement and test theory

There is ongoing debate as to what constitutes an educational assessment (James, 2010), how educational assessments should be structured (Mislevy, Steinberg, & Almond, 2003) and what they should be used for (Masters & Forster, 2000). The ever-changing nature of educational assessments and their co-existence with the available technology has also been acknowledged in the literature (Bennett, 2010). Educational assessment was identified as equivalent to the educational measurement in the four editions of flagship NCME and ACE publications (Brennan, 2006a; Linn, 1989; Thorndike, 1951; Thorndike, Angoff, Lindquist, & American Council on Education., 1971). However, the latest edition of the '*International Encyclopedia of Education*' gives educational assessment a more holistic meaning as stated by James (2010) "… educational measurement, in the form of tests, is often an element of educational assessment but the latter generally has a wider scope in terms of purpose, form, agency and use." (p. 161). Formative assessment is used mainly to generate feedback to learners in order to improve their performance (Black & Wiliam, 2009). This form of assessment does not require measurement. However, assessments that are designed to generate scores for individuals or higher level units (e.g. classes, schools or countries) do

require comparable metrics, and for this purpose, measurement is a requirement (Michell, 1997). PISA was conceived as a cross-national comparative study to facilitate policy development across countries and for this purpose, measurement is required. Numerous threats are identified to valid and reliable measurement (Wu, 2010), and violations of the assumption of LII are among them.

Different sources list the diverse types of functions that educational measurements serve. Linn (2010) asserts that educational measurement has four primary functions. Firstly, that educational measurement can be used in decisions regarding individual students (e.g., placement, high school graduation, or admission to university). Secondly, it functions as a tool for monitoring of education (e.g., through various sample-based comparative assessments). Thirdly, educational measurement can provide accountability measures (e.g., preferably based on whole population assessments). Lastly, it can serve a formative role for teachers and students to provide ongoing feedback. Black and Wiliam (2007) concurred with Linn's functions in part, accepting three purposes, those being certification, accountability and learning. A brief overview of the types of educational assessment offered by OECD (2009c, p. 94) simplifies this further describing two purposes of tests. In this publication, it states that on the one hand, educational assessments can be employed to measure an individual student's performance, in most cases for the purpose of selection or placement. Consequently, those evaluations are designed to maximise precision related to an individual examinee's estimated performance. On the other hand, OECD (2009c) points out that tests like large-scale international assessments are primarily designed to measure knowledge or skills of whole populations. In this case, minimising error on an individual level is of less concern, and effort is put into a reduction of the error while making conclusions about populations. This focus on the population's estimates is reflected in the need for a robust domain sampling that ensures that the test items administered provide a sufficient coverage of tested constructs (Crocker & Algina, 2008, p. 69). Multiple matrix sampling (MMS) designs (Beaton, 1987; Rutkowski, Gonzales, von Davier, & Zhou, 2013) are utilised in large-scale assessments to counterbalance an increased burden on test participants. In these sampling designs, each participant responds to a subset of items representing only part of the domain being tested, while ensuring that across the whole sample of examinees, all aspects of the domain are being assessed. Furthermore, the sample of examinees is reflective of the population. Thus, the population of interest is sampled, and the domain is fully represented. In addition, the burden on individual participants is reduced by an efficient test design in

which sets of items based on a common prompt enable greater coverage of the domain in a limited period of time. A disadvantage of this efficiency is the violation of the assumption of LII. While other aspects of measurement have been investigated in many other studies, the research reported in this thesis focuses on the extent of violations of the LII assumption and possible causes and effects of this violation.

The history of the first formalised and recorded performance assessments can be dated to about 200 B.C.E. when Chinese candidates for civil and military service needed to show sufficient skills, such as reciting passages for memory or marksmanship (Madaus & O'Dwyer, 1999). In the early 20th century modern test theory emerged with the establishment of Classical Test Theory (CTT) (Traub, 1997) and initial developments on techniques of item response modelling (nowadays referred to as the Item Response Theory or IRT) followed (Bock, 1997). Under CTT, each item is scored right or wrong, and items scores are summed to yield a total score. The observed total score is taken as an indicator of the student's true score, with the recognition that the observed score has an error component, although individual items do not. Item response theory assumes that some items are harder than others and that some test takers are more able than others. Under the original IRT models, the observed set of item scores for all test takers are used iteratively to estimate a difficulty parameter for each item and an ability parameter for each test taker. Each parameter is an estimate, and as such it has a margin of error. Crocker and Algina (2008) offer a comprehensive overview of both theories with brief introductions available in Cappelleri, Jason Lundy, and Hays (2014) and De Champlain (2010). IRT has gained popularity over the last few decades and branched out extensively to non-educational research fields, particularly in health (Hays, Morales, & Reise, 2000; Thomas, 2011). Despite the possible decline in the use of CTT in favour of IRT, there are still assessment situations in which CTT may be suitable (Zickar & Broadfoot, 2008). Over the last half of the century, IRT has been undergoing constant enhancement (Fayers, 2007). Unidimensional dichotomous data are typically modelled by using one, two or three parameter logistic models (Yen & Fitzpatrick, 2006). IRT models that can be utilised with unidimensional polytomous item data (Penfield, 2014) include the partial credit model (Masters, 1982), rating scale model (Andrich, 1978) and generalised partial credit model (Muraki, 1992). Several types of multidimensional models have been developed along with the more recent IRT techniques such as testlet, group-level or non-parametric models (Yen & Fitzpatrick, 2006). Comprehensive overviews of well established and emerging IRT models are offered by Reise and Revicki (2015) and

Linden (2016).

2.2 Brief overview of methodology used in scaling of PISA cognitive data

Although IRT models have been used frequently and successfully in many assessments and research projects, large-scale assessments require a more complex approach (Lietz, Cresswell, Rust, & Adams, 2017). The intention of these large-scale assessments is to collect comprehensive data about student achievement in different cognitive domains, such as mathematics, science, reading, problem-solving, and in sub-domains such as geometry and arithmetic. Consequently, to achieve the desired broad coverage of the domains assessed, and to limit the demand on students' time, the balanced incomplete block (BIB) design was implemented (Frey, Hartig, & Rupp, 2009; Gonzalez & Rutkowski, 2010; Mislevy, Beaton, Kaplan, & Sheehan, 1992; van der Linden, Veldkamp, & Carlson, 2004). Under this approach, instead of each participant responding to all possible items in the test, subsets of items are selected representing the main- and sub-domains of the test, and each student undertakes only a sub-set of all possible items.

PISA's cognitive data are scaled by applying a generalised form of Rasch model, i.e. mixed-coefficients multinomial logit model (MCML) (Adams et al., 1997; Adams & Wu, 2007). This model is used in three steps (OECD, 2009b). Firstly it is applied to national level data under the assumption that students have been sampled from a multivariate normal distribution. This step serves the purpose of evaluating the quality of data to identify items that have unacceptable psychometric characteristics and therefore may be excluded from remaining analyses. Secondly, this model is used with international calibration data (OECD, 2009b, p. 153) under the same assumption as national data calibrations. Thirdly, in a step of student scores generation, the MCML model is used again on "a country-by-country basis with the item parameters anchored at the values that were estimated in the international calibration" (Adams, Wu, & Carstensen, 2007, p. 276). The step is part of estimating the marginal posterior distributions in preparation for drawing of plausible values. The MCML model is likely to be also used in the evaluation of field trial outcomes when new items to be used in subsequent PISA studies are tested (OECD, 2008, 2009b).

The plausible values method was used for the first time in the USA in the National Assessment of Educational Progress (NEAP) (Beaton, 1987), while its theoretical

underpinning was given by Rubin (1987). A condensed overview of the PV methodology is given by Wu and Adams (2002) who regard plausible values as a representation of the range of abilities that a student might reasonably have. Many articles give a more detailed treatment of this methodology (Mislevy, 1991; Mislevy, Beaton, et al., 1992; Mislevy, Johnson, & Muraki, 1992; Monseur & Adams, 2009; von Davier, Gonzalez, & Mislevy, 2009; Wu, 2005). As this methodology is predominantly used in large-scale studies such as TIMSS or PISA, all technical or data analysis tutorials include brief information about it (Adams & Wu, 2002; Martin, Mullis, & Chrostowski, 2004; OECD, 2005a, 2005b, 2009b, 2012, 2014b). Usually, five plausible values are randomly drawn from the estimated distribution of students' abilities. This approach is in contrast to the estimation of the single point estimate of a student's ability under CTT and basic IRT models. However, it is still essential to obtain students' abilities from IRT models to impute plausible values.

2.3 Brief review of the literature on the merits of the PISA study

Since PISA's inception, a growing number of publications have pointed to various advantages and disadvantages of this large-scale assessment (Hopfenbeck et al., 2017).

Some critiques against PISA are related to largely non-psychometric aspects and include: (a) challenges related to possible linguistic and cultural biases (Arffman, 2010; Asil & Brown, 2016; Bonnet, 2002; Feniger & Lefstein, 2014), (b) concerns whether the PISA study measures what it is intended to measure (Bautier & Rayou, 2007; Le Hebel, Montpied, & Tiberghien, 2014), (c) uncertainties about the comparability of highly diverse educational systems (Berliner, 2015; Ercikan et al., 2015; Feniger & Lefstein, 2014; Prais, 2003, 2004), (d) utilisation of PISA for the sake of cross-national league tables (Kell & Kell, 2014b; Meyer & Benavot, 2013), and (e) philosophical arguments against the PISA study in line with general opposition to any standardised educational testing (d'Agnese, 2015; Gorur, 2011; Popkewitz, 2011; Serder & Ideland, 2016).

On the other hand there are also non-technical arguments supporting this initiative, arguing that PISA: (a) gives quality evidence for educational policy change (Carvalho & Costa, 2015; Schleicher, 2017), (b) provides data for subsequent educational research (Anderson et al., 2007; Domínguez et al., 2012; Kanes, Morgan, & Tsatsaroni, 2014; Sireci, 2015), (c) facilitates the comparison of its results with other studies (Brown, Micklewright, Schnepf, & Waldmann, 2007) and (d) offers frameworks supporting teaching (Ikeda, 2015).

Extensive debate has been conducted predominantly related to the technical and psychometric aspects of the study, discussing issues such as (a) choice of scaling model (Goldstein, 2004; Kreiner & Christensen, 2014; Oliveri & von Davier, 2014; Rutkowski, Rutkowski, & Zhou, 2016), (Kreiner, 2011) with a rebuttal by Adams (2011) (b) sampling procedures (Prais, 2003) with a rebuttal by Adams (2003) and rejoinder (Prais, 2004), (c) quality of cross-wave trends (Cosgrove & Cartwright, 2014; Jerrim, 2013; Wetzel & Carstensen, 2013) and (d) differential item functioning (Kreiner, 2011) with a rebuttal by Adams (2011). Many authors take a moderated approach to the technical challenges facing large scale educational assessments by acknowledging them, yet highlighting the strengths and complexities of undertaking international comparative educational studies (Schleicher, 2017; Shiel & Eivers, 2009; Wu, 2010). The foregoing discussion shows that PISA has clear strengths and identified technical challenges. The current research seeks to contribute to this literature by investigating the extent to which estimates of PISA scores may be compromised by the presence of LID.

2.4 Definitions and taxonomies of local item independence and their relations to other IRT assumptions.

According to Bock (1997), the principle of local independence in relation to Item Response Theory was introduced by Lazarsfeld (1950). As the whole report was initially requested by the American War Department during World War II, the definition of local independence presented there was narrowed down to the dichotomously classified responses of multiple-item attitude questionnaires. Lazarsfeld defined local independence as "… statistical independence of response from one item to another of persons having the same value as that of the underlying latent variable" (Bock, 1997, p. 24). However, the problem relating to the existence of inter-item dependencies and redundant measurement information was recognised even before IRT was clearly developed. Zenisky, Hambleton, and Sireci (2003) and Keller, Swaminathan, and Sireci (2003) indicated that a number of classical test theorists (Anastasi, 1961; Guilford, 1936; Kelley, 1924; Thorndike, 1951) had highlighted possible problems when items related to a common stimulus or scenario. A short historical background to LII is also offered by Henning (1989).

Local item independence is defined in two ways; a strong definition and a weak definition (Yen & Fitzpatrick, 2006, p. 122). According to Lord and Novick (1968, p. 361), for a group of examinees with the same value for a latent trait, LII is observed when "the conditional

distributions of the item scores are all independent of each other." This strong definition can be mathematically presented as:

$$P\left(U_1 = u_1, U_2 = u_2, \mathsf{K}\ , U_n = u_n \mid \theta\right) = \prod_{k=1}^{n} P\left(U_k = u_k \mid \theta\right) \quad (1)$$

where

$P()$ stands for the probability of an event given the ability in the parentheses,

U_i is the response variable for the *i*th item

McDonald (1979, 1981) proposed the weaker definition for which:

$$P\left(U_i = u_i, U_j = u_j \mid \theta\right) = P\left(U_i = u_i \mid \theta\right) \cdot P\left(U_j = u_j \mid \theta\right) \quad (2)$$

where

$$i, j \in \left\{1, 2, \mathsf{K}\ , n\right\}$$

While equation (1) in the strong definition relates to any subset of items, equation (2) from the weak definition has to be fulfilled only for pairs of items. Consequently, by using the weaker definition, the number of subsets of items that have to be tested is limited only to pairs, while in the case of strong definition all combinations (e.g. pairs, triplets, quadruples) of items have to be tested. Violation of local item independence is termed local item dependence although some authors prefer to use the term conditional dependence (De Ayala, 2009, p. 20) as being more descriptive.

Local item independence is one of three assumptions of unidimensional item response theory models. The second assumption is called monotonicity, and it means that there is a monotonic association between item performance and the probability of an examinee correctly responding to an item (Nandakumar & Ackerman, 2004). Some authors refer to monotonicity in a more general perspective as a functional form assumption (De Ayala, 2009, p. 21) requiring the data to trail the function imposed by the IRT model used, which frequently involves logistic or normal ogive functions. The third assumption is related to the number of dimensions under investigation, but in this research reduces to the unidimensionality assumption that

> states that the observations on the manifest variables (e.g., the items) are solely a function of a single continuous latent person variable. (De Ayala,

2009, p. 20)

However, as acknowledged by Yen and Fitzpatrick (2006, p. 122) dimensionality has been defined in different ways, and one of the dimensionality definitions proposed by McDonald (1981, 1982) links it explicitly to local item dependency. This causes some publications to regard LII and unidimensionality as equivalent which is not the case, as indicated by De Ayala (2009, p. 20), when items related to a common prompt or involved in item chaining may produce LID, yet not create a new dimension. The same author also gave an example of speediness when LID is present due to an additional dimension of students' rapidity not being modelled.

The two situations presented above are the backbone to, what Chen and Thissen (1997) proposed and that was supported by Houts and Edwards (2015), namely that the segregation of LID into "Surface LD" (SLD), i.e. driven by items' similar content or placement and "Underlying LD" (ULD) when a number of latent variables to model data are under factored.

Another categorisation of local item dependency was proposed by Marais and Andrich (2008a, 2008b) who called LID that was attributable to a common item stimulus "Trait dependence" whilst when a student response to an item is conditional upon a response to a preceding item, i.e. item chaining, then a "Response dependence" can be found (Andrich, 2016). These two types correspond to what Hoskens and De Boeck (1997) called "Combination dependence" and "Order dependence", respectively.

Yen (1993) in her highly cited LID literature paper mentioned another categorisation of LID, namely positive and negative LID, which reads as follows

> Positive LID means that, if a student performs higher (or lower) than expectation on one item, he or she probably will perform higher (or lower) than expectation on the other. The expectation is based on overall test performance. Negative LID means that, if a student performs unusually well on one item, he or she probably will perform unusually poorly on the other. (Yen, 1993, p. 188)

As an applied example of negative LID, Yen offered an example of the impact of time management when selective effort is given by the student to one section of the assessment as opposed to another section. Bolsinova and Tijmstra (2016) also discussed negative LID in light of response accuracy and response time interrelationships. The issue of negative LID is also discussed extensively by Habing and Roussos (2003) who argue that negative LID must

be present if positive LID is observed between the items in a testlet. An argument related to Habing and Roussos (2003) approach was also proposed by van Rijn and Rijmen (2015) who argued negative dependency due to "explaining-away phenomenon". Both the Yen (1993) and Habing and Roussos (2003) approaches to negative dependency appear to be somewhat in contrast with one another, with Yen arguing it to be driven by students' choice of selective time allocation and Habing and Roussos describing it as a mathematical artefact of within-testlet positive dependency.

Habing and Roussos (2003) mentioned that in the applied papers they reviewed, negative LID is very often ignored. Indeed many recent publications also ignore negative dependency altogether (Chen, Hwang, & Lin, 2013; Chung et al., 2014; Crins et al., 2016; Hissbach, Klusmann, & Hampe, 2011; Kalpakjian, Tulsky, Kisala, & Bombardier, 2015) or report it, yet do not attempt to explain its nature (Cole et al., 2005; Jones, Tommet, Ramirez, Jensen, & Teresi, 2016; Mattsson, Fearghal, Lajunen, Gormley, & Summala, 2015; Rodriguez & Crane, 2011; Williams et al., 2009). Yet, Braeken (2011, p. 62) suggested that negative local dependencies may be indicative of a general problem with the scale for which items do not measure something in common. Few publications offer explanations for negative LID that included guessing (Crane et al., 2012) or other reasons (DeMars, 2013; Kaspar & Hartig, 2016; van der Lans, van de Grift, & van Veen, 2017). Positive and negative LID will be investigated in this thesis.

2.5 Causes of local item dependence

Frequently, theoretical introductions to the local item dependence are followed by applied examples of LID pointing to common passage or other stimuli that is shared among a few items. However, this is only one of the many plausible drivers of LID. Yen (1993, p. 188) provided a comprehensive summary of possible causes of local item dependence, some of which may lead to substantial LID while others could yield negligible LID. The causes may relate to characteristics of items, examinees, and test administration. Listed below are ten causes for LID suggested by Yen, organised into three groups.

LID related to characteristics of items:

Item or response format

Similar designs of items may add an additional relationship between them. This refers to

students' skills in dealing with a particular item layout, and not purely with their performance on the trait being tested.

Passage dependence

Probably the most frequently recognised reason for LID occurs when a few items may be linked to a common reading passage, picture, graph, or other prompt. Students' special knowledge, or a lack of it, about a particular passage may unintentionally impact their responses to a collection of items linked to this passage.

Content, knowledge, and abilities

A collection of items that measure the same range of content (e.g. fractions in mathematics), or items with the same coverage in the curriculum or everyday life, can also be found to be dependent on each other.

Item chaining

Items can be organised in a way that a certain response to an item alters the chance of answering a subsequent item correctly.

Explanation of the previous answer

This cause of LID can be treated as a special case of item chaining for mathematical or scientific assessments when students are explicitly asked in one question to give an explanation of the reasoning behind the answer given to a previous question.

LID related to examinees:

Practice

Types of items that students have already been exposed to and have practised before the test administration may add dependence between them.

Fatigue

Groups of items at the end of the test may be dependent on each other as additional factors of fatigue, and lowered motivation may appear.

LID related to test administration:

Speediness

Items at the end of the test, for which students have not had enough time to respond, may lead to LID. Also, items that attract students' attention and to which students have allocated extra time can cause LID.

External assistance or inference

Students' performances can be unusually high for a group of items where external support of a teacher or fellow student was given. Similarly, interferences such as classroom disruptions or faulty materials can lead to a lower performance than would be expected under normal conditions for a group of items.

Scoring rubrics or raters

Using a common scoring rubric or the same raters for a set of items can lead to LID.

2.6 Negative implications of local item dependence

The presence of LID can have different implications on the various aspects of item response models. This section offers a brief overview of negative consequences of LID for IRT models.

Effects of LID on item parameter estimates

Estimates of item discriminations can be falsely inflated when LID is present (Ackerman, 1987; Ferrara, Huynh, & Baghi, 1997; Ferrara, Huynh, & Michaels, 1999; Reese, 1995; Sireci, Thissen, & Wainer, 1991; Thissen, Steinberg, & Mooney, 1989; Tuerlinckx & De Boeck, 2001a, 2001b; Wainer & Wang, 2000). This may have a negative impact when items are selected for a test construction or in computer adaptive testing. This is so because highly discriminating items are often preferred by test designers (Yen, 1993). Estimates of item difficulties may become more homogenous (Ackerman, 1987), or as some research suggested, underestimated, when the occurrence of LID is not taken into account (Jiao, Wang, & Kamata, 2005; Reese, 1995). The conclusions, mentioned above, were confirmed by Chen and Wang (2007) for positive LID, with the opposite effect of underestimation of the discrimination and overestimation for the difficulty parameters for negative dependency.

Ravand (2015) showed that the precision of item difficulties could be overestimated when LID is present. Estimates of item guessing parameters tend to be underestimated (Reese, 1995; Wainer & Wang, 2000).

Effects of LID on person ability estimates

Ackerman (1987) found that ability estimates were influenced as the degree of dependency increased. However, he recognised limitations of his results as local item dependence was simulated only for easier items. Spray and Ackerman (1987) found that a person's abilities were underestimated in the presence of LID. Reese (1995, 1999) further elaborated on this issue and discovered, with substantial levels of LID in the data, ability estimates can be overestimated in the low range and underestimated for the higher range of abilities. This was also confirmed in a study by Smith (2005), Marais and Andrich (2008a) and Ravand (2015). However, while Smith (2005) argued that person estimates are robust against the LID, research by Marais and Andrich (2008a) reached the opposite conclusion. A paper by Eckes (2014) showed that when ignoring testlet based dependency, the precision of person ability estimates is overestimated. This result was also confirmed by Ravand (2015).

Effects of LID on reliability, test information and standard error of measurement

LID can lead to the overestimated reliability and test information (Eckes, 2014; Keller et al., 2003; Reese, 1995; Sireci et al., 1991; Teker & Dogan, 2015; Thissen et al., 1989; Thompson & Pommerich, 1996; Wainer & Thissen, 1996; Wainer & Wang, 2000). The standard error of measurement (SEM) of trait estimates is the reciprocal of the square root of the test information. Therefore, when test information is overestimated, SEM is underestimated. As SEM and test information are used to construct confidence intervals for scores and to assess the number of items to achieve required accuracy (Yen, 1993), its misspecification can lead to negative consequences for validity and reliability of the students' scores. Also underestimated SEM may lead to negative consequences for computerised adaptive testing, resulting in a premature termination of the testing procedure (Reese, 1995).

Effects of LID on equating

The occurrence of LID may have an impact on equating procedures. This issue was initially indicated by Reese (1995). A paper by Monseur and Berezner (2007) showed with reading data from PISA 2000 and 2003 that the linking error could be underestimated due to local

item dependency arising from the use of testlets. A study by Cao, Lu, and Tao (2014) showed that a two-parameter logistic model was quite robust to LID for some equating methods investigated by them. On the contrary, other studies using three-parameter logistic models (Chen, 2014; Lee, Kolen, Frisbie, & Ankenmann, 2001; Zhang, 2007) suggested that LID has a negative impact on equating. Because of the confounding effect of LID on equating, Hastedt and Desa (2015) explicitly excluded data from PISA study from their simulation study, investigating the relationship between a number of items used and linkage error.

Other negative effects of LID

While the majority of publications listed in this section focus on unidimensional models, a simulation study by Brandt (2012) looked at the effect of LID on multidimensional IRT models. His conclusions suggested that covariance estimates in multidimensional calibrations could be biased by testlets instigating local dependence.

Consequences of local dependence in relation to computer adaptive testing (CAT) was evaluated by Pommerich and Segall (2008) and Walter and Rose (2013). Pommerich and Segall (2008) found that LID has a limited effect on the precision of examinee CAT scores, but they questioned their approach to simulating LID. Walter and Rose (2013) looked at LID driven by the item ordering and concluded that this type of dependency has little effect on item calibrations and person ability estimations. Frey and Seitz (2011) evaluated the usefulness of multidimensional adaptive testing (MAT) with PISA data, pointing to the various advantages of MAT, compared to a non-adaptive approach. While no dependency was investigated, the same authors (Frey & Seitz, 2009) identified item dependence as potentially problematic for MAT.

Teker and Dogan (2015) investigated local item dependence in relation to differential item functioning (DIF). They concluded that LID could inflate the number of items exhibiting DIF as well as overestimate its levels.

2.7 Strategies for managing local item dependence

Yen (1993) highlighted six procedures for either reducing LID or analysing data taking LID into account, in order to reduce possible adverse consequences for the validity of obtained measurements. Those procedures can be executed at different stages of test construction and data analysis. Firstly, before a test is administered to students, extra care can

be taken to construct independent items. Secondly, the test should be administered under appropriate conditions so as to minimise possible sources of LID, for example, speediness, fatigue or inappropriate interference or assistance. Thirdly, scoring for items that are locally dependent on each other can be combined. However, there are limitations attached to this particular strategy as "… it may be difficult to determine *a priori* which responses are locally dependent [and] … it can take additional rater time and decrease the quality of the ratings if a scoring rubric becomes too complicated." (Yen, 1993, p. 210). Fourthly, items can be reviewed for LID after a test has been conducted. With the added assistance of LID indices, the process of reviewing would be far more efficient for subsequent applications of the test. Fifthly, LID can be dealt with by constructing separate scales, one with items that would be expected to be highly dependent on each other, and the second scale with the remaining items. However, Yen (1993) indicated that this procedure does not eliminate the need for an accounting of LID in those separate scales. Finally, using a testlet strategy was suggested by Yen (1993) and by Wainer and Kiely (1987). In this procedure, dependent items are, at the stage of calibration of the scale, combined into one partial credit item. Using this strategy may facilitate using tests with an authentic design that purposely incorporates dependent items. However, this strategy also leads to loss of information, as fewer items are used for calibration of the scale (Keller et al., 2003).

There is a large number of different IRT models proposed to accommodate LID such as: Bayesian based models (Almond, Mulder, Hemat, & Yan, 2009; Bradlow, Wainer, & Wang, 1999), Hierarchical generalized linear model (Jiao, Kamata, Wang, & Jin, 2012; Jiao et al., 2005), Boundary mixture model (Braeken, 2011), Rasch subdimension model (Brandt, 2008, 2010) or a variety of other testlets based models (Chen, 2012; Huang & Wang, 2013; Ip, 2010; Paek, Yon, Wilson, & Kang, 2009; Wainer & Wang, 2000; Wilson & Adams, 1995), also reviewed in Wainer, Bradlow, and Wang (2007).

2.8 Detecting the local item dependence

2.8.1 Classification of LID detection methods

This section offers a classification of the LID detection methods, leading to a more detailed overview of a few selected indices starting with the index used in this study. Many methods have been suggested for detecting local item dependence. These methods generate indices that can be classified in various ways. One way could be according to the type of definition of the local item independence that is of interest. Two definitions of LII are

recognised in the literature (see point 2.4), the strong and the more commonly used weak definition. In most practical applications, the strong definition has not been used. The vast majority of methods of LID detection take as a basis the weak definition, either in its standard version that focuses on pairs of items, or less stringent requirement that the conditional covariance between items is zero. The latter of those two is frequently used in the context of nonparametric IRT. That leads to a second way by which methods of LID detection can be classified, according to the type of Item Response Modelling used. A third way of categorising methods of LID detection emphasises whether or not LID stems from additional dimensions to those that are required by the particular IRT model being implemented. Finally, another categorisation depends on the items that can be investigated with the particular LID indices, as some can be applied only to dichotomous items, others to polytomous items, or to both.

A comprehensive taxonomy of pairwise item indices was given in Kim, deAyala and Ferdous (2011) who reported work undertaken by Kim (2007) in his PhD that reported 16 LID indexes, as reproduced in Table 2.1.1. He separated them into a three-dimensional classification that took into account three characteristics of those indexes, namely:

> (1) whether they require item or person or both parameter estimates in calculation (IRT-based vs. non-IRT-based), (2) whether they are significance tests or measures of association (significance test vs. measure of association), and (3) whether they can detect the direction of item dependence (e.g., positive LID and negative LID) (directional vs. non-directional) (Kim, 2007, p. 14)

Table 2.1.1. Table adapted from (Kim, 2007, p.14) providing three-way classification of pairwise LID indices

Residual Correlation in Factor Analysis	Non-IRT-based	Directional index	Measure of Association
Modification Index in Structural Equation Modeling	Non-IRT-based	Non-directional index	Significance Test
Mantel-Haenszel Estimator for Common Odds Ratio Mutual Information Difference Q_3			Measure of Association
Cochran-Mantel-Haenszel Test Standardized Log-Odds Ratio Difference (SLORD) Standardized Phi (φ) Coefficient Difference (SPCD). Fisher's r-to-z Transformed Q_3 Wald Test from logistic regression	IRT-based	Directional index	Significance Test
Absolute Mutual Information Difference			Measure of Association
Glas's Modification Index Likelihood Ratio G^2 Likelihood Ratio Test from logistic regression Pearson's X^2 Power-Divergence Statistic	IRT-based	Non-directional index	Significance Test

Kim's (2007) work seems to be the most comprehensive overview of methods for detecting LID. However, his research focused mainly on pairwise item indexes. Other methods for quantifying LID are reported. Ferrara and Huynh focused on identifying LID based on raw scores through the assessment of the magnitude of the inter-item correlations across groups of examinees with similar raw scores (Ferrara et al., 1997; Ferrara et al., 1999; Huynh & Ferrara, 1994; Huynh, Michaels, & Ferrara, 1995). Another approach for detecting LID was suggested by Sireci et al. (1991) and Wainer and Thissen (1996). These authors suggested that in cases where it is possible to recognise a testlet like the structure of the test with, for example, a number of reading items clearly associated with particular reading passages, then reliability estimates of such a test can be obtained in two ways. In the first, the apparent testlet structure could be ignored and the reliability of the test calculated, with all items assumed to be independent. A second way would take the testlet structure into account and then again calculate reliability estimates. If LID is present, reliability estimates for the second method would be lower than the first one. This approach, however, gives a rather rough indication of LID.

Douglas, Kim, Habing, and Gao (1998) also focused on detecting LID between pairs of items. However, they used conditional covariance curves that were "… estimated by a simple extension of kernel smoothed IRF [item response functions] and scaled to account for variation in the item difficulty" (Douglas et al., 1998, p. 148). The main difference between this method and the majority of LID indices reported by Kim (2007) comes from the possibility of looking at how the conditional covariance of two items varies across the range of the students' ability distribution. This approach may allow recognising not only the presence of LID but also shed some light on the reasons why LID occurred. This method aims to test for LID defined under a less stringent definition than that which requires the conditional covariance between items to be zero.

New approaches for detecting LID have emerged (Adams & Wu, 2009; Debelak & Arendasy, 2012; Kreiner & Christensen, 2013b; Liu & Maydeu-Olivares, 2013; Liu & Thissen, 2012) with particular focus on the use of multidimensional item response models (Bao, Gotwals, & Mislevy, 2006; Bartolucci, 2007; Wang, Cheng, & Wilson, 2005). Posterior predictive model checking was proposed in research led by Levy (2011; 2009) with the same author extending earlier work to propose Generalized Dimensionality Discrepancy Measure (Levy & Svetina, 2011; Levy, Xu, Yel, & Svetina, 2015) as a new tool for detection of the local item dependency. Methods of testing for LID in the graphical loglinear Rasch models (Kreiner & Christensen, 2004, 2011) or graded response models (Liu & Thissen, 2014) also developed. More work has been dedicated to assessing LID in polytomous and dichotomous scored items (Ip, 2001; Tsai, Chaimongkol, & Hsu, 2006) and to incorporating additional information about response-time (van der Linden & Glas, 2010) while

assessing dependence. Finally, additional developments occurred in refining the definition and detection of the response-dependence (Andrich, Humphry, & Marais, 2012; Marais & Andrich, 2008b) and utilising an estimate of the change in the item difficulty because of its dependence on another item as a tool for quantifying the magnitude of dependence between two items (Andrich & Kreiner, 2010).

2.8.2 Detailed overview of three indices for detecting LID

Although there are many techniques for detecting LID, quite often the proposed methods are used and reported with real data only in the research papers that introduced them. However, some indices, such as Q_3 statistic, introduced by Yen (1984), or residual correlation in factor analysis, appear to be the more popular and widely accepted methods for LID detection. Therefore they are reviewed in more detail below. Also the modification index in CFA models is described in more detail. All three LID indices were recommended by Kim (2007) and Kim et al. (2011) who found that:

> The power of the Fisher's r-to-z transformed Q_3 was the highest among the 10 LID indexes across all the LID conditions. …There was no LID index that performed the best with respect to all three aspects: Type I error rate, power and false positive rate. However, Q_3, MID, residual correlation, AMID and the Modification Index could be recommended for most of the LID conditions because of their relatively high power and low false positive rate. Kim (2007, p. 4)

Residual Correlation in Factor Analysis

Analysis of the residual correlation matrix in factor analysis can be used for testing whether LII is violated. If LID is not present in the data, then only one factor related to an underlying student's ability trait should be observed in the data. Therefore, after fitting a single-factor model to the data, the examination of a residual correlation matrix should show that the values in this matrix are close to zero. If this is not the case for some pairs of items, this indicates that LID is present. The justification for the residual correlation (RC) index to be chosen for this study is presented in section 3.3.3.1 of the methodology chapter, while the delimitation about the size of the cut point is discussed in section 3.3.3.2. In this section, the RC index definition and its utilisation for LID detection with a particular focus on challenges related to the interpretation of negative residual correlation are outlined.

The existence of positive residuals implies that the model under-predicts the covariance between two variables whereas negative residuals indicate that model predicted covariance is too high (Bollen, 1989, p. 257). Looking at this definition from a different perspective as stated by Maydeu-Olivares:

> A positive residual correlation implies a model expected correlation [smaller][2] than the observed correlation, whereas a negative residual correlation implies an observed correlation [smaller] that the model expected correlation." (Maydeu-Olivares, 2015, p. 118)

Using residual correlations from factor analyses as a LID assessment approach has been utilised in a number of applied papers with an educational background (Hissbach et al., 2011; van der Lans et al., 2017). Notably this approach is more popular in research related to health sciences, and it was used while researching item banks (Amtmann et al., 2010; Haley et al., 2009; Kim, Chung, Amtmann, Revicki, & Cook, 2013; Reeve et al., 2007), implementing computer adaptive testing (Flens et al., 2017; Resnik, Tian, Ni, & Jette, 2012; Smits, Zitman, Cuijpers, den Hollander-Gijsman, & Carlier, 2012) and investigating the quality of various measures (Anatchkova et al., 2014; Cook et al., 2007; Crane et al., 2012; DeMars, 2013; Jones et al., 2016; Watt et al., 2014).

While local dependence due to a common stimulus is likely to be represented by high positive RCs for pairs of items from the within same testlets, other sources of LID may be indicated by high values of RCs from pairs of items across different testlets. Approaches to the interpretation of negative residual correlations in a factor analysis setting used in scientific articles vary. Some authors ignore negative RCs altogether. Hissbach et al. (2011) only reported absolute values of residual correlations. DeMars (2013, p. 187) also ignored negative residual correlations, and stated that their presence was "… likely as an artefact due to the small number of items on each scale and the inclusion of the items score in both the observed and predicted score." Cole et al. (2005) acknowledged the existence of a large negative residual correlation but labelled their interpretation as "elusive" and did not pursue the issue. Similarly, Rodriguez and Crane (2011), Mattsson et al. (2015), and Jones et al. (2016) found a negative residual correlation between item pairs while investigating the factorial structure of their measures but offered no practical explanation as to the nature of this negative RC.

Three other applied papers mentioned negative residual correlations in relation to factor analyses and offered some practical explanations for their presence. In the first paper, Crane et al. (2012) found that two items from a word recognition task within the Alzheimer's Disease Assessment Scale, that aimed to measure memory, showed negative residual correlation. These two items were part of the same assessment task with one aimed at quantifying true positive responses and the other one true negatives. The authors proposed that this negative RC reflected the guessing strategies that their elderly participants applied to this task. Salonen et al. (2017) in their paper investigating dimensionality of the tool for evaluating gambling, suggested that bias in reporting

[2] The original text states "larger" which is incorrrect. Personal communication with the author confirmed the mistake.

could be responsible for negative residual correlation between a question about hiding the evidence of gambling and admitting to having a betting problem, with participants who cover up the addiction less likely to concede it. In the third paper related to the educational domain and discussing the development of an instrument for teacher feedback, van der Lans et al. (2017) found two items showing negative RC. While they offered a plausible explanation for it, they acknowledged that its presence required additional investigations. The limited literature located and reported above points to challenges in acknowledging and explaining negative residual correlations.

Modification Index in Structural Equation Modelling

The modification index (Sörbom, 1989), equivalent to the Lagrange multiplier test, is commonly used for model modification in structural equation modelling (Kaplan, 1995). Steinberg and Thissen (1996, p. 92) mentioned that one of the reviewers of their article quoted "… After fitting a one-factor model to data, the modification indices … for the error covariance matrix could be examined, with large values taken as indicating LID". Kim (2007) used this idea and incorporated this index into his study, finding it to be a valuable means for testing for the presence of LID. This index was used as a LID indicator in an applied paper by Hill et al. (2007) or another LID indices evaluation by Houts and Edwards (2015). While the structural equation modelling (SEM) literature sometimes reports that MIs of value 3.84 is indicative of a non-negligible improvement in the fit of the model (Brown, 2015; Teo, 2013), this cut point is not universally accepted. For example, the Mplus software uses the value of 10 as the default for MIs which are reported in the outputs. Finally, literature also points to the importance of allowing for the correlated residuals which are theoretically design driven regardless of suggestions offered by MIs (Westfall, Henning, & Howell, 2012). This index is discussed here as it was used in this research (see section 5.4.1).

Yen's Q_3

While this index is not utilised in this study, it is discussed here as it is a widely used pairwise LID index belonging to the IRT class of indices as per classification proposed by Kim et al. (2011). This index is also used in one other publication (Monseur et al., 2011) purposely dedicated to LID investigation in PISA that is reviewed in detail later in this chapter. Finally, a recent publication by Christensen, Makransky, and Horton (2017) provided important results about this index that were factored while addressing research aim number three.

The Q_3 statistic, introduced by Yen (1984), for dichotomous item response models has been the most popular index for the detection of Local Item Dependence. When it was introduced, another

available statistic called Q_2 (van der Wollenberg, 1982) was found to have various shortcomings. Consequently, Yen (1984, p. 127) proposed Q_3 as an alternative for measuring local dependence. In the thirty years since its introduction, this index has been used in many research studies not exclusively related to education (Ackerman, 1987; Balazs & De Boeck, 2006; Buysse et al., 2010; Chen & Thissen, 1997; Huynh et al., 1995; Ip, 2001; Keller et al., 2003; Kim, 2007; Lee, 2004; Pommerich & Segall, 2008; Skaggs, 2007; Yen, 1993; Zenisky, Hambleton, & Sireci, 2002; Zenisky et al., 2003).

The Q_3 statistic is defined as the correlation between the residual scores for two items i, j, computed across all examinees. Therefore,

$Q_3 = r_{d_i d_j}$ where $d_{ik} = u_{ik} - \hat{P}_i\left(\hat{\theta}_k\right)$ is the residual score for *kth* examinee calculated as the difference the between the observed score for *i* item and *kth* examinee, and corresponding probabilities of a correct response based upon the estimated item and ability parameters from the applied latent trait model.

Although the Q_3 statistic has been used frequently, most researchers reported its descriptive properties by reporting the distribution of Q_3 (Chen & Thissen, 1997; Yen, 1993) or by checking the mean values for item pairs for either whole tests (Ackerman, 1987; Balazs & De Boeck, 2006; Keller et al., 2003), or subgroups of items that were of interest for those researchers (Huynh et al., 1995; Lee, 2004; Skaggs, 2007; Zenisky et al., 2003). At the same time, some concerns related to this statistic were raised by Chen and Thissen (1997) or by Houts and Edwards (2015), as well as suggestions for more detailed studies that could evaluate the performance of Q_3 (Kim, 2007). Despite those concerns about the Q_3 statistic, many researchers have used it in their work and adopted a practical cut point of .2 for identifying the existence of LID. An important advancement in this regard was proposed by Christensen et al. (2017) who argued for the Q_3 cut points values needing to be related to the number of items involved in the test as well as the sample size. Another more recent development in regard to this index was proposed by Finch and Jeffers (2016) who proposed a Q_3-based permutation test for LID. The Q_3 index is discussed in here as these recent developments are likely to further extend the popularity of this approach to LID detection and is also mentioned in the final chapter regarding suggestions for future research.

2.9 Reported occurrences of local item dependence in educational studies and PISA in particular

Local item dependence has been investigated on many occasions by researchers who used

different datasets from either smaller or non-international school data based studies (Allen & Sudweeks, 2001; Bao et al., 2006; Ferrara et al., 1997; Ferrara et al., 1999; Gustafsson & Rosen, 2006; Ip, 2001) or educational but medically related datasets (Lawson & Brailovsky, 2006; Zenisky et al., 2002).

Literature discussing local item dependence in PISA studies is scarce, with only ten publications located that mention LID in conjunction with PISA data. However, only one of them by Monseur et al. (2011) is purposely dedicated to LID investigations. Five papers (Cai, 2010; Cai, Yang, & Hansen, 2011; DeMars, 2006; Koğar & Kelecioğlu, 2017; Trendtel et al., 2014) are devoted predominately to proposing and evaluating various psychometric models (largely IRT) and use limited parts of the PISA data in applied examples, pointing to the existence of LID. Two research investigations by Kreiner (2011), and mostly overlapping the 2011 report, a paper by Kreiner and Christensen (2014) discuss various psychometrical challenges of PISA discussing LID as one of them. Finally, three papers investigate language invariance (Oliden & Lizaso, 2013), and item-position effects (Debeer & Janssen, 2013) with item dependence being a part of the discussion.

All the publications listed in the previous paragraph use only selected subsets of PISA data, with Cai et al. (2011) working with only six mathematics testlets from PISA 2000 and students data limited to the single booklet from five nations participating in the study. A paper by Cai (2010) selected only 15 mathematics and 32 reading cognitive items using a random sample of PISA 2000 single booklet users. This data choice was matched closely by DeMars (2006). The single booklet was also used by Koğar and Kelecioğlu (2017) who took advantage of mathematics data from PISA 2012. Reading data from PISA 2009 was employed in investigations by Trendtel et al. (2014) and by Oliden and Lizaso (2013) who used national level data from Germany and Spain, respectively. Both publications by Kreiner (2011) and Kreiner and Christensen (2014) utilised booklet 6 with reading items from PISA 2006. The report by Monseur et al. (2011), explicitly dedicated to LID in PISA, focused on the mathematics dataset from PISA 2003 and a reading dataset from PISA 2000. A paper by Debeer and Janssen (2013) also utilised PISA 2006 data from all three cognitive domains but is limited to Turkey's dataset.

Nine out of the ten publications focused predominantly on LID related to the existence of testlets, with an only paper by Debeer and Janssen (2013) looking at the contribution of the location of the items within the assessment, i.e. item-position effects towards item dependency. Among these nine papers, five publications (Cai, 2010; Cai et al., 2011; DeMars, 2006; Koğar & Kelecioğlu, 2017; Trendtel et al., 2014) gave overall evidence for LID due to testlets by showing that psychometric models factoring testlets fit better. The publication by Trendtel et al. (2014)

looked in detail at only one testlet with the remaining publications (Kreiner, 2011; Kreiner & Christensen, 2014; Monseur et al., 2011) offering more details about the testlets investigated in their reports. Detailed results from these publications involving specific testlets are incorporated in the summary sections of the results (see Section 5.4.1.2 and 5.4.1).

2.10 Summary

In summary, this literature review aims to provide a concise overview of educational measurement leading towards an introduction into the methodology used in scaling the PISA cognitive data. A perspective debating the strengths and weaknesses of the PISA study is also offered. The definition of local item independence and different classifications of the violation of this IRT assumption are discussed. Different drivers that may induce LID are discussed as some of them will be hypothesised to explain the results observed in this research. To highlight the importance of the LII assumption, the negative implications of item dependency are reviewed, followed by a short discussion about the ways in which LID could be managed. A brief review of methods of detecting LID is offered to focus on the approach used in this project. The chapter concludes with the review of the limited literature which reports the presence of local item dependence in PISA.

An extensive literature search failed to find a publication that gives an overview of testlets used in PISA from the perspective of cross-wave usage, for the purpose of the equating of PISA studies. This gap in the literature is proposed to be addressed by the first research aim of this study.

The reviewed literature reporting LID in PISA is scarce and limited to small subsets of the PISA data without examining cross-wave consistency or the comparing of LID prevalence across cognitive domains or types of item dependency. This is of interest in the second research aim. This research aim also incorporates investigations into different plausible LID drivers aiming to extend the published research which mostly targets LID due to testlets.

None of the located publications mentioning LID in PISA investigates whether the existence of LID is more or less prevalent in some countries participating in the PISA assessments, nor do they consider differential testlet functioning. These literature shortcomings will be addressed as part of research aim number three investigations.

The next chapter proposes detailed research questions for each of the general research aims proposed in the introduction to this thesis. It also discusses analytical plans for addressing these research questions along with discussing approaches to the data collections and delimitations.

CHAPTER 3 METHODOLOGY

The current study investigates the prevalence and likely causes of item dependency in the PISA study, a large-scale international and longitudinal assessment program. Because of this focus, it is necessary to use secondary data as it would not be feasible within a PhD research project to design and develop large-scale assessments, nor would it have the authenticity of the original source data for the specified research objective. The main purpose of the PISA study is to provide comparative data on the performance of national, and in some countries, subnational, education systems. The use of PISA data for the current purpose renders this a secondary analysis. According to Vartanian (2011), secondary datasets have several advantages, e.g. the cost and time involved in designing, developing and administering instruments, the ability to follow units (in this case countries) over time. In this research, the aim is to consider sets of test items that are used longitudinally, to build upon the high quality of the data given the investment of resources and international expertise in the development of the instruments, the sampling design, quality assurance processes over sampling, item development and testing, and survey administration. While these advantages are substantial, secondary data analysis has some disadvantages; the most substantial for the current study are the lack of information about some items and the inability of the researcher to conduct follow-up interviews with participants in order to understand why, for example, they may have skipped items. Some items are released and are therefore available for inspection, but other items are kept secure by the survey managers as those items are scheduled to be re-used in future assessments. For non-released items, limited information is available, so inferences about possible causes of item dependency are constrained. A further consequence of the use of secondary data is that decisions about the research design are restricted. The research must, therefore, operate under the ontological and epistemological positions that guided the development of the PISA study. The current research, therefore, takes a pragmatic stance to the research problem (Punch, 2014).

As a consequence of this position, the research design is an empirical, but not an experimental, one. In addition, the research is exploratory in that it seeks to reveal the extent of local item dependency and to identify possible causes of it. While it has been customary to ascribe LID to the use of common item stems (Wainer et al., 2007), the current research takes a more open stance and does not propose hypotheses to explain LID but seeks to exploit all available information about test items and to build explanations for the observed LID from that information.

3.1 List of detailed research questions

Although the general aims and research objectives were introduced in the introductory chapter,

this section expands them to offer more detailed research questions corresponding to each objective.

Research questions for research aim 1 - Description of PISA's testlets

The first research aim is to describe the testlets used in the five PISA waves and three cognitive domains tested in the study. Below, two detailed research questions are addressed in Chapter 4.

RQ_1A - What are the longitudinal patterns of testlet usage across five waves of PISA (from 2000 to 2012)?

RQ_1B - How do testlets vary in regard to the within-testlet variability of item difficulty estimates?

Research questions for research aim 2 – LID in data from PISA's international calibrations

The second research aim is to investigate the existence of LID in data from PISA's international calibration datasets and to offer explanations for LID. Four detailed research questions corresponding to research objectives flagged in Chapter 1 are proposed and answered in Chapter 5.

RQ_2A - What is the prevalence of positive and negative LID in the international calibration data?

RQ_2B – To what extent does the prevalence of LID vary by cognitive domain, PISA wave or within-testlet or between-testlet location of pairs of items?

RQ_2C - What evidence is available to support possible explanations for any observed LID, whether it be positive or negative?

RQ_2D - Which testlets show cross-wave consistency in the presence of local item dependency?

Research questions for research aim 3 - LID in data from PISA's national calibrations

The final research aim intends to investigate the existence of LID in data from PISA's national calibrations from 24 OECD countries. The four detailed research questions matching research objectives flagged in introductory chapter are listed below. They will be addressed in Chapter 6.

RQ_3A – To what extent does the prevalence of LID vary between countries?

RQ_3B - If countries show increased levels of local item dependence, what factors might explain this?

RQ_3C - Is the presence of LID in national calibration level data consistent when compared to international calibrations?

RQ_3D - Which testlets suggest a presence of differential testlet functioning between countries on the basis of LID presence?

3.2 Methodology for research aim 1 - Description of PISA's testlets

3.2.1 Plan for the data analysis and software

In order to address research questions for this research aim, a number of items within each testlet are pictorially represented with the help of high-low-close charts which feature maximum-minimum-average values of the percentage of correct item responses. The graphs were panelled by PISA wave to facilitate across time overview. IBM SPSS Statistics software (IBM Corp., 2015) was predominantly used for analyses addressing this research aim.

3.2.2 Data preparation

As this chapter aims to provide a descriptive overview of the testlets used in the PISA studies without entering into a debate about the advantages and disadvantages of various mathematical models that might be used for scaling cognitive data, the percentage of correct responses for each cognitive item is reported as it is the simplest indicator of aggregated item difficulty.

To undertake analyses, the percentage of correct responses based on international calibration datasets for each cognitive item and from five PISA waves (2000, 2003, 2006, 2009 and 2012) were obtained. These data were extracted from tables reporting main study item pool classifications that are provided in the appendices of technical manuals (Adams & Wu, 2002; OECD, 2005b, 2009b, 2012, 2014b).

3.3 Methodology for research aim 2 - LID in data from PISA's international calibrations

3.3.1 Plan for the data analysis and software

The cognitive data from five PISA studies and three domains were used in order to conduct fifteen Confirmatory Factor Analyses from which the databases with residual correlations were extracted. Detailed explanations about the data preparation procedures are offered below in Section 3.3.2. The choice of residual correlation from factor analyses as a LID indicator is elaborated in a discussion of the delimitations of this study (see Section 3.3.3.1). The CFAs were estimated using weighted least squares using the mean and variance adjusted (WLSMV) estimator which is suitable for categorical data and for which robustness has been tested (DiStefano & Morgan, 2014) and recommended (Finney & DiStefano, 2013). The Mplus software (Muthén & Muthén, 2015) was employed. For the domains which were a primary focus in each assessment occasion (reading in

2000 and 2009, mathematics in 2003 and 2012 and science in 2006) five analyses were conducted with Mplus but with the help of high-performance computing (HPC) Tizard machine services offered by eResearch SA (eResearch South Australia, 2017). This was necessary as the evaluations of each PISA wave primary focus cognitive domains, on average, utilised over 110 items. Trial runs showed that each country's estimation for these domains took about 6-8 hrs with a standard university computer. While the Mplus licence allowed the use of only one HPC processor at the time, taking advantage of automation of jobs within the HPC streamlined the estimations particularly for national level data.

The research aim utilising data from PISA international calibrations had four subsequent research questions which were addressed by the following methods.

Research question 2A, aiming to estimate the prevalence of positive and negative local item dependency, was addressed by reporting the percentages of residual correlations exceeding a threshold proposed in section 3.3.3.1 as indicative of local item dependency. In order to combine the results across five PISA waves, a meta-analytical approach was undertaken with the help of Comprehensive Meta-Analysis software (Borenstein, Hedges, Higgins, & Rothstein, 2014). A random effects model (Borenstein, Hedges, Higgins, & Rothstein, 2010) was applied for each cognitive domain to combine LID prevalence results across five PISA implementations and to calculate meta-analytical 95% confidence intervals. The same approach was used while looking at the within-testlet and between-testlet location of item pairs while addressing research question 2B.

Investigations related to explaining the reasons for observed dependency, thus addressing research question 2C, implement a dual approach. Firstly, a qualitative analysis is undertaken which takes advantage of PISA items that are released to the public. Positive and negative dependency is visualised by utilising network graphs available in the R package *qgraph* (Epskamp, Cramer, Waldorp, Schmittmann, & Borsboom, 2012) used in a few applied research papers (Kuittinen et al., 2017; Mattsson et al., 2015). The qualitative conclusions are also supported by the availability of electronic appendices reproducing the released items to facilitate quick access to these cognitive questions.

Secondly, quantitative analyses aimed to investigate LID drivers are reported. This is done in Section 5.4.2 of Chapter 5 through reporting multilevel hierarchical logistic regressions using LID presence as the outcome variable and item pair characteristics, e.g. difference in item difficulty, as explanatory variables in the models. The level 1 data used for the analyses represent pairs of cognitive items for which RCs were obtained from the Mplus based CFAs. As some items were

used multiple times across the PISA waves, the same pair of items could be featured in the dataset up to five times. The multilevel models were used to control for the nesting effect of an item pair. Therefore, the item clusters that formed the Level 2 observations were established from the pairs of the PISA item identifiers. In total each cognitive domain has separate models predicting positive LID events as expressed by RCs greater than or equal to 0.1 and negative LID events identified when RCs are less than or equal to -0.1. In total, six models are reported with the reference category for all of them being "Trivial RC" which represents RC between -0.1 and 0.1.

The procedure *melogit* (StataCorp, 2015) from Stata 14.2 (StataCorp, 2017) was used to estimate the odds ratios with a random intercept model being used (Liu, 2016, p. 380). The interaction terms were not incorporated into the models because, despite overall large sample sizes, a number of events of interest were relatively small, raising concerns about assumptions regarding a number of events per estimated parameter (Peduzzi, Concato, Kemper, Holford, & Feinstein, 1996) and the "zero cells" problem (Hosmer & Lemeshow, 2000, p. 135). The limitation of the small number of item pairs in each nested group was investigated. On the basis of suggestions by Moineddin, Matheson, and Glazier (2007), it is expected that with a large number of groups present, the parameter estimates and their SEs should be recovered correctly despite small group sizes. The analyses were undertaken in a hierarchical way starting with the null model followed by the inclusion of independent variables that are constant across all cognitive domains, such as PISA study year, whether the pair of items is located within the same testlet or cluster. In subsequent models, predictors were added related to the item source, the language in which the item was developed, and the difference in difficulties of item pairs. Finally, variables representing item properties unique to the cognitive domain were added. The final models reported in the text were selected following recommendations by Kuha (2004) who suggested that both AIC and BIC fit criteria should be considered along with significance tests.

In the case of modelling negative LID, a perfect separation issue resurfaced. For the reading and science domains, there was not a single event of negative LID when pairs of items were located in the same testlet. Allison (2008) offered a number of suggestions for single-level logistic regressions pointing to penalised likelihood estimation methods as the best solution. However, no recommendation was made in his paper with regards to multilevel analyses. Ensoy, Rakhmawati, Faes, and Aerts (2015) only mention a two-stage Bayesian-based approach (Abrahantes & Aerts, 2012) as a plausible robust approach for clustered data. As this method has not been widely cited and evaluated as yet, the decision was made to stay with the default melogit procedure of dealing with the perfect separation that removes the cases from the level of the variable producing complete separation. This renders the negative LID models to be estimated on subsets of data excluding item

pairs from the same testlet. In the mathematics domain, there was a single data point with negative LID, so the melogit did not implement case removal. However, to retain consistency with the other two domains this negative LID within-testlet single item pair was removed.

Research question 2D, which looks at cross-wave consistency in the LID presence, is addressed in section 5.4.1 along with the qualitative investigations of the positive and negative dependency.

IBM SPSS Statistics (IBM Corp., 2015) was used for data management, descriptive statistics and all other analyses not mentioned above. In order to address any subsequent data analysis queries, the research and analyses diary has been kept by utilising IBM SPSS syntax and Stata Do-files.

3.3.2 Data preparation

This section elaborates all the steps undertaken to produce the dataset used in addressing research questions featured in Chapter 5.

3.3.2.1 Preparation of the PISA datasets

Scored cognitive datasets from PISA 2006, 2009 and 2012, which are openly available, were used[3]. In the case of datasets from PISA 2000 and 2003, an additional data preparation step had to be undertaken, as only students' unscored cognitive responses were available in the online data collection (OECD, 2015e). The scoring for these two early PISA waves followed the credit allocation rules provided in cognitive codebooks available online (OECD, 2015a). Item-level nonresponse, as well as multiple or invalid responses, were scored as "No credit". This follows the approach used by OECD in scoring, for example, PISA 2009 cognitive data. Following the rule reported in PISA technical manuals, not-reached items were treated as 'not administered'. Furthermore, students from schools who catered for students with special needs (i.e. those who responded to the UH booklet) were excluded from the analyses. A similar approach to the data from this booklet has been exercised for the majority of the analyses undertaken by the PISA team. Following the PISA's approach to national calibrations (Adams & Wu, 2002, p. 101; OECD, 2009b, p. 146) unweighted data was used for both results chapters.

The international calibration cognitive responses dataset was prepared in the following manner. The first set of 15 datasets, reflecting five waves and three cognitive domains, incorporated

[3] PISA 2000, 2003, 2006 and 2009 data used throughout the research were downloaded on 15th of Jan 2015 while PISA 2012 data were obtained on 12th of Dec 2016. Consequently, any changes made by PISA to the cognitive datasets after these dates are not applied in this research.

subsamples of 500 students from only 26 OECD countries[4]. Only countries which participated in international calibration samples for all five waves were incorporated. As an exception to the rule above, data from Mexico were not used, as this country opted for a different set of easier cognitive items in the last two waves of PISA, as compared to the 26 chosen countries. This way of preparing international sample calibration datasets maintained the consistency of selected countries across the five waves as well as the coherence of the same cognitive items used by all countries in each PISA wave[5]. The simple random subsamples of 500 students from each country's datasets were stratified by different strata variables following the procedures indicated in the PISA Technical Manuals (OECD, 2014b, p.163).

Consideration was given to an alternative approach to preparing international level data for this study to precisely mirror the procedures used by PISA teams in the preparation of international calibration samples following guidelines in the PISA technical manuals for each wave. This was challenging for the international calibration datasets for PISA 2009 and 2012. A description of the sample selection process for international calibration in the corresponding technical manuals was somewhat unclear. The PISA technical team from ACER that conducted the study was approached, seeking clarification about the creation of the international calibration samples. Feedback from Berezner and Timms (2016) confirmed a few minor typing errors in PISA 2009 and 2012 technical manuals, but also clarified changes to international calibration data organisation compared to the initial two waves. For the sake of cross-wave consistency, the researcher made the decision to work with 26 OECD countries only.

3.3.2.2 Preparation of the Mplus input files

As elaborated in the previous section, Mplus software (Muthén & Muthén, 2015) was used to conduct Confirmatory Factor Analyses (CFA) in order to address research aims two and three. While for PISA 2003-2012 the preparation of Mplus input files and selection of cognitive items to be used was straightforward, PISA 2000 required additional steps. In PISA 2000 and in contrast to all the later waves, an unbalanced booklet design was used. This resulted in some pairs of cognitive items never being responded to by the same student, and therefore covariance estimates for such pairs could not be obtained. For example, a cluster[6] of science items called S4 was allocated to Booklet 4 and Booklet 9. However neither of these two booklets incorporated items from cluster S1.

[4] Australia, Austria, Belgium, Canada, Switzerland, Czech Republic, Germany, Denmark, Spain, Finland, France, United Kingdom, Greece, Hungary, Ireland, Iceland, Italy, Japan, Korea, Luxembourg, Norway, New Zealand, Poland, Portugal, Sweden, United States

[5] Reading cognitive data from the USA for PISA 2006 is not available as due to booklets' printing error the PISA consortium made decision not to use this data.

[6] In PISA 2000 a term "Block" was used which in all later PISA studies converted into "Cluster".

To obtain the estimates of residual correlations for the majority of pairs of PISA 2000 science items, the CFA was run twice. Initially, clusters S1, S2, S3 were used and on the second occasion items from clusters S2, S3, S4 were employed. Values of residual correlations for items from S1 and S4 came from separate CFAs, and RCs for items from S2 and S3 were averaged from both runs. As an additional step, an intraclass correlation coefficient (ICC) using the ICCVAR procedure (Hedberg, 2016; Hedges, Hedberg, & Kuyper, 2012) was calculated to justify aggregating two RC estimates for items from clusters S2 and S3. For example, ICC and 95%CI for science was 0.96 (0.95, 0.97) indicating that within run RC estimates were very similar despite being extracted from different CFAs. Because of the issue mentioned above, 64 estimates for residual correlations involving items from different testlets could not be produced as compared to what would be possible if a fully balanced booklet design was also applied for PISA 2000. The same approach was used with PISA 2000 mathematics[7]. PISA 2000 reading required three runs of CFAs with different settings: (1) items from clusters R1 to R7 were used, (2) items from clusters R1, R3, R7, R8 were used, and (3) items from clusters R8 and R9 were employed[8].

3.3.2.3 Preparation of the residual correlations datasets

As Mplus produces output results in text format limited in width to 90 characters, dedicated IBM SPSS syntax to extract residual correlations was used. As this syntax has a potential to be reused more widely, it is made available as an Electronic Appendix 3.3.1. Each row in the final RC dataset reports the RC value for a specific pair of items for international calibration samples and each country involved. The information from the PISA Technical Manuals about items source, the language of submission, cluster and difficulty level expressed as international percent correct, was merged into the final RCs datasets. This was extended by merging other characteristics of the cognitive items. The section below reports in more detail the steps and challenges involved in the process of locating information about the items.

3.3.2.4 Search for information about cognitive items

In order to provide plausible explanations for observed LID a considerable amount of background data preparation was required.

Firstly, an extensive online search was undertaken to locate the PISA items that have been

[7] ICC and its 95%CI for mathematics was 0.96 (0.95, 0.97). Eighty one residual correlations involving items from different testlets could not be estimated when compared to the fully balanced booklet design implemented in later waves of PISA.
[8] ICC and its 95%CI for reading was 0.95 (0.94, 0.96). Just over 2600 residual correlations involving items from different testlets could not be estimated when compared to the fully balanced booklet design implemented in later waves of PISA.

released. OECD PISA established a website (OECD, 2016b) as well as a similar NCES site (National Center for Education Statistics, 2016b) dedicated to providing released PISA items. The publications listed on both web pages provided the majority of released items, but some information came from less apparent sources (OECD, 2009a). Table 3.3.1 shows that of the 179 testlets utilised in this research, about 39% have been located as released testlets. The prevalence of released items varies by cognitive domain with close to 50% of testlets from mathematics being available for in-depth interpretation.

Table 3.3.1 Proportions of PISA testlets released to the public[9]

		Cognitive Domain							
		Mathematics		**Reading**		**Science**		**Total**	
		n	%	n	%	n	%	n	%
IS THE TESTLET RELEASED?	**No**	41	52%	37	67%	32	71%	110	61%
	Yes	38	48%	18	33%	13	29%	69	39%
	Total	79	100%	55	100%	45	100%	179	100%

Secondly, each PISA's wave has a publication elaborating the assessment frameworks for the cognitive domains (OECD, 1999, 2004, 2006, 2010a, 2013). Each framework provides details on various item dimensions that were investigated. For example in PISA 2000 mathematics items were allocated by (a) Overarching Concept also called "main mathematical theme" (Growth and change, Space and shape), (b) Item Type (multiple choice, closed constructed-response, open-constructed response), (c) Item context (Community, Educational, Occupational, Personal, Scientific), (d) Item Competency Class (Class 1: reproduction, definitions, and computations / Class 2: connections and integration for problem solving / Class 3: mathematical thinking, generalisation and insight) and finally (e) Mathematical Content Strands (Algebra, Functions, Geometry, Measurement, Number, Statistics). Obtaining information about the item dimensions was driven by a desire to identify plausible causes of LID for pairs of items which were not released to the public. Although in PISA 2012 some information on the items' characteristics is available in the Technical Manual, this is not the case for any previous PISA waves. A large number of sources were used to collate additional information about items' characteristics. Most sources are OECD publications such as reports, technical manuals, assessment frameworks, and information accompanying released items. The search was not limited to these sources. For example, the majority of the information about mathematical items came from a single English language table located in a French language

[9] The single item testlets are also counted in this table

publication (DEPP, 2007, p. 149). Similarly, a considerable proportion of science item characteristics were extracted with the help of Google Translator from a Czech language report (Mandíková & Bašátková, 2008). Information about the characteristics of reading items were derived from somewhat secondary sources including a French publication (Soussi, Broi, Moreau, & Wirthner, 2004) and a report produced by a Finnish university (Sulkunen, 2007).

On occasions, for released items which had no other sources for item characteristics, the researcher's own judgement was exercised in coding the items following definitions from PISA's frameworks. If the information about the item was located in references pointing to a specific PISA wave, that item description was propagated to all the waves which also used this item. The types and naming of item dimensions of interest occasionally vary across PISA waves as driven by changes to the assessment frameworks for different PISA waves. However, a common 'across waves coding' has been proposed. For example, items labelled in PISA 2000 as "Short response", were deemed to be related in PISA 2003, 2006 and 2009 to a "Closed constructed response" type, which in turn for PISA 2012 approximated a new type labelled "Constructed Response Manual". For the purpose of analyses in Chapter 5, all of these different naming conventions were recorded into "Short response". Possible drawbacks of such cross-wave standardising are acknowledged in the limitations section (section 7.2.1).

The efforts of finding information regarding items resulted in very few items' characteristics being missing. For reading, all information about items' aspect (e.g. "Access and retrieve"), text type (e.g. "Argumentative", "Chart/Graph"), situation (e.g. "Occupational", "Personal"), and text format (e.g. "Continuous", "Non-continuous") was located. Also science resulted in 100% data saturation for items' application (e.g. "Frontiers", "Health"), context (e.g. "Personal", "Global"), content (e.g. "Knowledge of science - Physical systems", "Knowledge about science - Scientific enquiry") and competencies (e.g. "Using scientific evidence", "Explaining phenomena scientifically"). Searching for properties of mathematics items was fully successful for items' situational placement (e.g. "Scientific", "Public") and content (e.g. "Space and shape", "Change and relationships"). For 4% of questions, the type of mathematical process (e.g. "Employ", "Interpret") involved could not be located. The missing data rate for mathematical competency (e.g. "Connections", "Reproduction") was 20%. Two additional items' characteristics: item length (e.g. "Long", "Medium") and mathematical strand (e.g. "Geometry", "Algebra") located in the literature resulted in 80% data availability. Furthermore, for all three cognitive domains, 100% of item information was located in regard to item format (e.g. "Complex Multiple Choice", "Short response"), item difficulty and language of submission to the PISA study. All the items' characteristics were merged into data with the residual correlations obtained from Mplus

estimations.

Cross-validation of extracted items' characteristics was undertaken by comparing located distributions of items' dimensions against published sources. For example looking at the published cross-tabulations for mathematics items in PISA 2003 (OECD, 2005b, p. 28-29), 2006 (OECD, 2009b, p. 45-46), 2009 (OECD, 2012, p. 43) showing the distributions of mathematics items according to item format, item content category and item competency concurred perfectly with data extracted in this study. Similarly, extracted item characteristics were verified against published tables (OECD, 2014a, p. 2-3) with the item distributions for PISA 2012.

Identifying and cross-validation of the mathematical item characteristics agreements proved to be more challenging for PISA 2000 as item characteristics were not reported consistently in the technical manuals. For example, two OECD publications elaborating on PISA 2000 results reported inconsistent tables (OECD, 2001, p.240) versus (OECD & UNESCO Institute for Statistics, 2003, p.266). A third source (Adams & Wu, 2002, p. 28) had to be located to address this inconsistency, but item formats in this publication did not agree with those reported in the PISA 2000 codebook. Another example comes from the PISA 2006 Technical Manual (OECD, 2009b) which reports on page 46 item counts that do not match item classifications on page 381. Similarly, the PISA 2003 Technical Manual misclassifies one item M413Q02 in two different parts of the same publication (OECD, 2005b, p. 257 and p. 412). Scale allocation for a number of items (for example M155Q04T) did not match across two PISA technical manuals (OECD, 2009b, p. 381) and (OECD, 2012, p. 336) and the third source needed to be consulted. It appears that Table A1.2 in the PISA 2006 Technical Manual misreported a considerable number of items' scale allocations. Similarly, disagreements have been detected for numerous reading items. For example item R406Q01 from the "Kokeshi Dolls" testlet was labelled by the PISA 2012 Technical Manual (OECD, 2014b, p. 410) as being of "Constructed Response Expert" item format, while the 2006 Technical Manual (OECD, 2007b, p. 15) suggests that this item is "Simple Multiple Choice". Other resources were consulted (OECD, 2015d, p. 23; Soussi, Broi, Moreau, & Wirthner, 2013, p. 108) to resolve these inconsistencies and generate final item format allocations. Similar item type disagreements between the sources, mentioned above, related to R412Q08 from "World Languages" and the item reading aspect for R452Q03 from "The Play's the Thing". Discrepancies mentioned in this paragraph and other similar ones were solved in favour of more official publications such as OECD produced technical manuals. This process of validation of item characteristics resulted in each of the item characteristics being assigned characteristics based on the citation from which the most consistent information has been obtained. This paragraph highlights challenges in locating information crucial to addressing one of the research questions but also suggest a possibility of erroneous item

characteristic allocation for some cognitive questions.

3.3.3 Delimitations

This section aims to introduce and provide scientifically based justifications for research choices and assumptions that have been made while addressing research aim number two regarding LID detection in PISA international calibrations.

3.3.3.1 Arguments for selection of non-IRT based LID index

The decision to use non-IRT based indices for detecting Local Item Dependency (LID) was taken for several reasons.

Firstly, publications by Kim and others (Kim, 2007; Kim et al., 2011) indicated that out of 10 LID indices investigated in their Monte Carlo simulation study, residual correlation (RC) from factor analysis was one of two indices that offered a good balance between maximum power and low false positive rates.

Secondly, it was decided to give a perspective on LID presence in PISA studies without entering into a contested debate, acknowledged in the literature (Kreiner, 2011; Kreiner & Christensen, 2014; Rutkowski & Rutkowski, 2016) versus (Adams, 2011; Berezner & Adams, 2017), about the type of IRT models that should be used in PISA. Furthermore, few publications (Erosheva, Fienberg, & Junker, 2002; Manrique-Vallier & Fienberg, 2008) suggest that Rasch model cannot capture negative dependence.

Thirdly, as mentioned in the literature review this non-IRT index was used in various applied papers investigating the presence of LID (Amtmann et al., 2010; Anatchkova et al., 2014; Cook et al., 2007; Crane et al., 2012; DeMars, 2013; Flens et al., 2017; Haley et al., 2009; Hissbach et al., 2011; Jones et al., 2016; Kim et al., 2013; Reeve et al., 2007; Resnik et al., 2012; Smits et al., 2012; van der Lans et al., 2017; Watt et al., 2014). The value of non-IRT index was also acknowledged in simulation study by Houts and Edwards (2015) who stated that

> Because polychoric correlations are obtained from raw data, and do not consider the model used to generate the data, they are available when assessing both SLD [surface local dependence] and ULD [underlying local dependence due to unmodeled latent variables] and provide a common metric on which to compare the two types of LD. (Houts & Edwards, 2015, p. 296)

Fourthly, the relation between categorical data CFA and IRT is well documented and commonalities highlighted. For example, research addressing the relationship between different forms of factor analyses and IRT models goes back to the mathematically proven equivalence of

marginal likelihood of the two-parameter normal ogive model in item response theory and factor analysis of dichotomized variables (Takane & de Leeuw, 1987). The validity of this point and Takane and de Leeuw (1987) contribution is acknowledged by Wirth and Edwards (2007). Kamata (2008) investigated this in detail using different parameterisations of binary factor analysis models and provided mathematical formulas for transforming FA parameters into IRT parameters. Kreiner and Christensen (2013a) argued that unidimensional CFA and IRT models have more resemblances than distinctions and the difference between both models relates mostly to different focuses when the model fit is investigated. Furthermore, they argue that

> It is our point of view that both types of analyses are valid and meaningful. To us, there is therefore no reason why CFA should not consider the fit of the item distributions to the CFA model and there is no reason why the IRT analyses should not be concerned about the fit of the marginal item correlations to the expectations of the IRT or Rasch model. (Kreiner & Christensen, 2013a, p. 2)

Smith (1996) empirically investigated correspondence of dimensionality estimations from unidimensional Rasch models and various factor or principal component analyses and found that they yielded comparable results. Other publications investigated the attitudinal Likert-style type of the data. Waugh and Chapman (2005) found the principal component analysis was not performing well, although the post-PCA residuals were not investigated by the authors. DeMars (2013) looked at CFA and the multidimensional Rasch partial credit model and argued that both approaches resulted in comparable assessments of dimensionality. A similar endorsement of CFA, used for the purpose of LID assessment, was given by Hambleton and Swaminathan (1985, p. 24) and by Linacre (2009).

Interestingly, despite both techniques (categorical data CFA and IRT) being in use for a lengthly time, the debate about using both or either for the sake of assumption testing seems to be an ongoing and current issue. This deliberation is best expressed in "Ask the Experts: Rasch vs Factor Analysis" forum in the journal of Rasch Measurement Transactions (Christensen, 2012; Engelhard, 2012; Salzberger, 2012). Out of three forum panellists, Salzberger (2012) sees little value in conducting FA to accompany Rasch modelling while Engelhard (2012) and Christensen (2012) suggest that FA could be used as a supplementary tool for model fit. A similar opinion was put forward by Wirth and Edwards (2007). Maydeu-Olivares, Cai, and Hernández (2011) looking specifically into the comparability of fit between FA and CFA models arguing that, for binary data, both models produce similar fits. The debate around this issue was likely the reason for a dedicated major review of this topic aimed at ordered categorical data which has been undertaken in the dissertation by Kappenburg -ten Holt (2014). The author reviewed 28 simulation studies that compared FA and IRT spanning the years 1985 to 2012. This review provided a set of expectations,

which in turn were investigated through Monte Carlo simulations. FA of the estimated polychoric correlation matrix using mean-and-variance adjusted weighted least squares was one of four models compared. It was found to perform well in regard to the estimation of item parameters and their corresponding standard errors, model fit indices, and latent variable scores as long as the latent variable of interest was normal regardless of item distributions. While in the case of a simulated skewed latent variable the authors preferred the IRT model, they also recommend that the FA polychoric approach could also be utilised because of its useful fit statistics.

The approach in which CFA serves as an additional tool for evaluating fit has been utilised in a number of applied papers (Hill et al., 2007; Snowden, Watson, Stenhouse, & Hale, 2015), while Randall and Engelhard Jr (2010) proposed to use it while investigating measurement invariance. A more detailed account of preference for FA, IRT or both being used in applied papers is given by Holt, van Duijn, and Boomsma (2010), who found that 78% of the studies they investigated used only FA techniques for scale construction and evaluation while an additional 7% implemented both. Critique of using CFA was reported by Tate (2003), and Stone and Yeh (2006). The possible impact of guessing on dimensionality, and possibly LID detection was indicated in their papers. While both papers stated that CFA performed comparably to other methods of dimensionality assessment, the papers also highlighted that most techniques investigated by them might be impacted by guessing. However, given that the IRT model used in PISA does not account for students' guessing, implementing approaches that can more robustly accommodate guessing would undermine the relevance of the results of the current study to scaling currently used in PISA.

In conclusion, it appears that there is no uniform agreement as to whether categorical data CFA (CCFA) should be used as an extra tool along with IRT for fit evaluation in educational assessments. However, recent recommendations (Barendse, Oort, & Timmerman, 2015) appear to lean towards the utility of CCFA, particularly for the sake of fit evaluation. Mathematical and simulation literature, briefly reviewed above, points to the comparability of the WLSMV estimator based CFA and IRT.

3.3.3.2 Discussion regarding the size of the residual correlation's cut off value

Different publications appear to be using different cut-points for residual correlation purported to flag LID. These publications use different primary sources to justify their choice, yet many such standard bearing sources do not provide a scientific justification for their recommendations. For example, Reeve (2007) proposed the use of an RC of 0.2 but did so without justification. His suggestions have been cited subsequently in other papers (Haley et al., 2009). Highly cited books by Kline (2013) as well as Kline (2016) argue that special attention needs to be given, while evaluating

dimensionality, to residual correlations exceeding the absolute value of 0.1. Few applied papers (Amtmann et al., 2010; Cook et al., 2007; Watt et al., 2014) follows this recommendation in regard to LID investigation. This threshold residual correlation value is used as an indication of local item dependency.

At the same time, Kline (2016) acknowledged that there is an inherent problem with such cut points as accurate indicators of type or degree of model misspecification. Recently new research has emerged, published by Christensen et al. (2017), which proved for another LID index that no single value of Q_3 statistic should be used as its null distributions are related to the number of items, number of response categories and the sample size. The authors proposed a bootstrapping procedure, involving in their case 10000 simulated datasets, to prepare empirical distributions of Q_3 statistic. The Christensen et al. (2017) approach was not exercised in this study due to time consuming estimations making it implausible to use bootstrapping. Also the size and complexity of the data used in this part-time study made it impossible to change the approach late in the candidature. The limitations related to using fixed cut-point are acknowledged in the section 7.2.1. Furthermore, the final results chapter involving national level datasets with largely varying sample sizes considered Christensen et al. (2017) results.

3.3.3.3 First order of CFA versus second order CFA for main cognitive domains.

Different cognitive domains were targeted in each iteration of the PISA study. In 2000 and 2009 the focal domain tested was reading, in 2003 and 2012 it was mathematics and in 2006 science was the key focus. PISA technical manuals explicitly list various subdomains of interest for each focal domain being investigated. For example in 2012 mathematics was examined with a particular focus on four sub-domains targeting different types of item content, namely: *Change and Relationships, Quantity, Space and Shape, Uncertainty and Data*. At the same time, the cognitive processes required while addressing the mathematics items were targeted with three processes categories of *Employ, Formulate and Interpret* involved. Furthermore, items were prepared to serve content and process evaluations simultaneously.

All categorical confirmatory factor analyses were undertaken as first order CFAs. While it could be argued that second order CFAs would be perhaps more suitable for targeted cognitive domains, this approach is in-keeping with the method for which scores for main domains were produced for the purpose of PISA reports. The PISA 2012 technical manual explains that a number of different IRT scaling models have been used for different scale generation purposes. In conjunction with the argument for the equivalence of categorical data CFA and IRT presented above, the second sentence in the quote provides a justification for using first order CFAs in this study.

> Five multi-dimensional scaling models were used in the PISA 2012 Main Survey. The first model, made up of one reading, one science and one mathematics dimension, was used for reporting overall scores for reading, science and mathematics. A second model, made up of one science, one reading and four mathematics scales, was used to generate scores for the four mathematics subscales Change and Relationships, Quantity, Space and Shape, Uncertainty and Data. A third model, made up of one science, one reading and three mathematics scales was used to generate scores for the three mathematics subscales: Employ, Formulate and Interpret. A fourth model, made up of one reading, one science, one mathematics, one digital reading dimension, one digital mathematics and one digital problem solving dimension was used for reporting overall scores for reading, science, mathematics and computer-based mathematics, digital reading and computer problem solving scales for countries that implemented the computer-based assessment (CBA) in the PISA 2012 Main Survey. A fifth model, made up of one reading, one science, one mathematics and one digital problem solving dimension was used for reporting overall scores for reading, science, mathematics and computer problem solving scales for those countries that implemented problem solving in the PISA 2012 Main Survey as the only computer-based component.
>
> (OECD, 2014b, p. 396)

Similarly, PISA 2009 and 2006 also used four (OECD, 2012, p .152) and two (OECD, 2009b, p. 146) multi-dimensional scaling models, respectively. In both these studies, the first models reported indicate that overall dimensions of mathematics, science, and reading are of interest. The first impression in the technical manuals for PISA 2000 and PISA 2003 appears to point to a somewhat different approach:

> The PISA model is five-dimensional, made up of three reading, one science, and one mathematics dimension
>
> (Adams & Wu, 2002, p. 101).

> In PISA 2003 the main scaling model was seven-dimensional, made up of one reading, one science, one problem solving and four mathematics dimensions.
>
> (OECD, 2005b, p. 122)

However, upon more detailed reading of both technical manuals, it is likely that IRT models including all mathematics items were used in PISA 2003 (OECD, 2005b, p. 187, and p. 256), which is also indicated by reporting plausible values for “combined mathematics scale” (OECD, 2005b, p. 130). Relevant sections from a technical manual for PISA 2000 also points to similar conclusions (Adams & Wu, 2002, p. 107) when the combined reading scale is mentioned.

Most likely two multi-dimensional scaling models were used in PISA 2000, and PISA 2003 in the manner of later waves but only the 2006 PISA Technical Manuals made it explicit. At least it can be argued from this, that an IRT model with all mathematical items has been used in PISA 2000 and

2003, and that could justify application of first order CFA also for PISA 2000 reading and PISA 2003 mathematics LID investigations.

3.3.3.4 Using two multilevel logistic regressions instead of multinomial multilevel logistic regression

In relation to quantitative analyses reported in section 5.4.2, a consideration was given to utilising multilevel multinomial logistic regression instead of two separate analyses for positive and negative LID. A number of arguments outweighed this approach in favour of using two triplets of separate models – one for positive and the other for negative LID. Firstly, as reported in section 5.4.2, multilevel logistic regressions are built and presented in hierarchical fashion leading to final multivariate models. Incorporating the multinomial analyses would render reporting of the results from many models even more complicated. Secondly, there seems to be no literature which would strongly argue for multiple logistic regression rather than two separate logistic regressions provided the reference category is common (Grace-Martin, 2017) with the exception that the former may have less statistical power (Agresti, 2002, p. 273-274). Given large sample sizes involved in the analyses, this should have limited consequences for the conclusions. Thirdly, conducting the multinomial logistic regression requires assessing the assumption of the independence of irrelevant alternatives (Kwak & Clayton-Matthews, 2002, p. 405). This assumption would require conducting two separate logistic regressions and compare their coefficients with the full multinomial model. Lastly, for one of the few relevant independent variables, namely identifying whether the item pair is in the same testlet, no cases exist for negative LID instances and this might lead to convergence problems for the estimations of multiple multilevel logistic regressions.

3.4 Methodology for research aim 3 - LID in data from PISA's national calibrations

3.4.1 Plan for the data analysis and software

In order to address research question RQ_3A involving the prevalence of LID for 24[10] OECD countries, an approach for flagging LID used for international data level was maintained. The results were descriptive in nature and visualised in graphs featuring all countries and PISA waves. Separate figures were produced for positive and negative dependency featuring in case of positive LID sub-separation into within-testlet and between-testlet location of items' pairs. IBM SPSS Statistics (IBM Corp., 2015) was used for data management and preparing of the graphs.

The findings from RQ_3A suggested the relationship of LID prevalence, i.e. size of the residual

[10] Two countries were excluded from this part of the analyses, and the reasons are provided in section 7.2.2.

correlations cut-point, on sample sizes of students' cohorts. This association is also featured in a recent publication by Christensen et al. (2017). The considerable varying sample sizes of national cohorts of students imposed changes to a methodological approach of addressing research question number RQ_3B, aimed at finding countries that may present higher levels of dependency. From a simulational study by Christensen et al. (2017), it appeared that all figures reported by the authors, which were used for suggesting sample size adjusted LID indicators, follow a reciprocal function. In this study, this function was fitted to the LID prevalence results from 24 countries. The model residuals were then investigated with the intention of identifying countries with outlying residuals. Fit curve function in statistical package NCSS (NCSS LLC, 2017) was used as it allows for bootstrapped prediction limits to be pictured. Many different approaches for identifying outliers are available (Aguinis, Gottfredson, & Joo, 2013). In this research, featured in NCSS, Grubbs' test (1950) for detecting outliers was used, along with Rosner's (2011) test for many outliers and a classical boxplot based approach (Tukey, 1977) was also employed. Boxplots featuring 24 data points were produced with Medcalc (MedCalc Software BVBA, 2017).

The LID cut-point comparability problem also weighted on a methodological approach to addressing RQ_3C and RQ_3D aimed to look at the consistency of dependence for international calibration data as well as across national results. Fractional ranks of residual correlations expressed as a percentage were calculated for each of 360 primary CFAs analyses. Section 6.3 is descriptive in nature featuring the tables with the selected pairs of items within-testlet and between-testlet looking at their cross-wave and cross-national consistency or its lack thereof.

3.4.2 Data preparation

The procedure for preparing datasets followed the same steps that were introduced in section 3.3.2. However, for this research aim, 24 national datasets and three cognitive domains were used for each of five PISA waves, resulting in 360 CFAs analyses. All residual correlations obtained from Mplus software were, subseqeuntly, merged into a single data file featuring close to 800,000 data points. As mentioned in the previous section, the comparison of national and international LID existence was made on the basis of fractional ranks expressed as a percentage which were calculated independently for each combination of PISA wave, cognitive domain and 24 OECD countries. Fractional ranks approaching 100% are suggestive of higher positive residual correlations while fractional ranks close to 0% point to considerable negative residual correlations.

CHAPTER 4 RESULTS FOR RESEARCH AIM 1 - DESCRIPTION OF PISA'S TESTLETS

4.1 Introduction

The Programme for International Student Assessment (PISA) is an international large scale educational survey involving large proportion of world's economies. It evaluates education systems by testing 15-year-old students' scientific, reading and mathematical literacy skills. Furthermore, the PISA study collects additional student level information such as home and family background. It also gathers data about schools from students' perspectives and directly by assessing various aspects of schools' organisation (Turner & Adams, 2007). Since the first data collection took place in the year 2000, there have been to date[11] six waves of PISA studies undertaken, each with a different major domain as the focus. Technical manuals (Adams & Wu, 2002; OECD, 2005b, 2009b, 2012, 2014b) of the first five waves give detailed accounts of steps undertaken at the test design and development stages.

Table 4.1.1 reports[12] the number of cognitive items administered across all waves and domains. The table excludes items for which PISA technical manuals did not provide information about their item parameters. The largest counts indicate the domain that was a primary focus in each assessment.

Table 4.1.1 Number of items used in five waves of the PISA study across three main cognitive domains.

	Year Tested				
	2000	2003	2006	2009	2012
Items used in reading assessment	**129***	28	28	**131**	44
Items used in mathematics assessment	31	**84**	48	35	**109**
Items used in science assessment	34	34	**103**	53	53

* Bold font indicates the cognitive domains which were targeted in each PISA wave.

While test design and data collection procedures may vary slightly from one PISA wave to another, the study aims to assess students' cognitive skills in reading, science and mathematics during two-hour tests. The majority of assessments and participating countries utilise a paper and pencil format with provision for digital assessment emerging as an elective option from the 2009

[11] As of 2017.
[12] Items removed from the final international calibrations are not included.

wave onwards. The PISA study focuses on assessing students' abilities to apply knowledge in real-life situations. This approach leads to utilising groups of questions for which a common introduction, an information graph or another stimulus is used for a collection of connected items. Following the suggestion of Wainer and Kiely (1987) the term used for such groups of items is a "testlet". PISA technical manuals use the term "unit" to refer to a group of items with a common stimulus, but this term is not widely utilised elsewhere. To avoid confusion, the term testlet is used throughout this thesis. Items belonging to the same PISA testlet can be identified by an overlapping label and testlet title. Figure 4.1.1[13], reproduced from PISA's publication providing sample of the released items (OECD, 2002), is an example of a testlet labelled R040 and called "Lake Chad". This testlet is made of five items that share a common stimulus and introduction. For the sake of consistency, situations in which an introduction, graph, or other stimulus material is followed by only one question, are called a single-item testlet. So far the PISA studies used testlets that ranged from one to seven items for reading and from one to a maximum of four questions for mathematics and science. Given limited assessment time, students did not respond to all cognitive questions but were randomly allocated to one of the booklets (nine used in 2000 with 13 implemented for later waves[14]). For most assessments, each booklet was composed of four clusters (each representing 30 minutes of test time) of reading, mathematics and science questions following principles of a balanced incomplete block design (van der Linden et al., 2004). Each cluster, in turn, incorporates few testlets.

[13] The version used in PISA did not include the information about the questions' characteristics or give answers.
[14] In PISA 2009 and 2012, two sets of 13 booklets were available, called "Standard Booklet Set" and "Easier Booklet Set".

READING UNIT 1

Lake Chad

Figure A shows changing levels of Lake Chad, in Saharan North Africa. Lake Chad disappeared completely in about 20000 BC, during the last Ice Age. In about 11000 BC it reappeared. Today, its level is about the same as it was in AD 1000.

Figure A

Lake Chad: changing levels

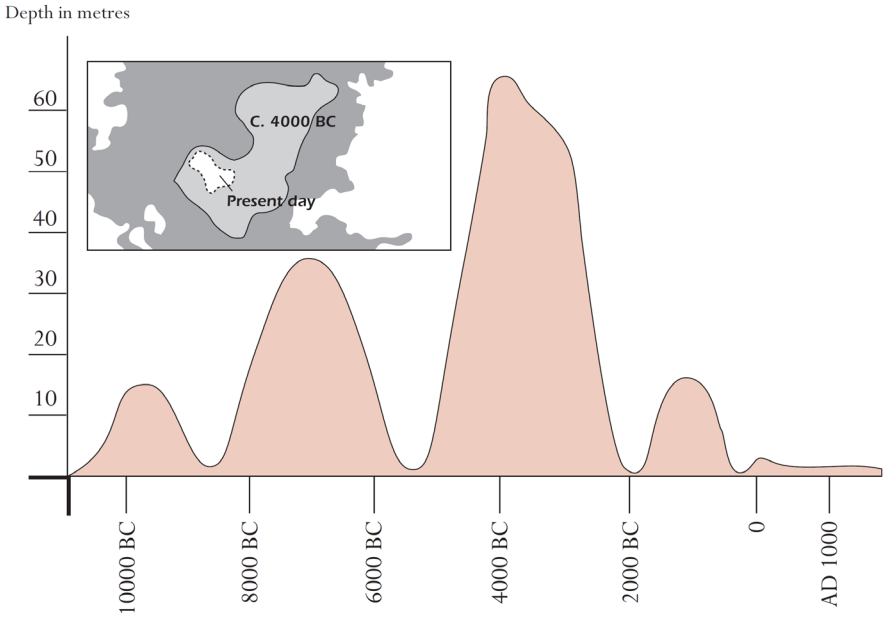
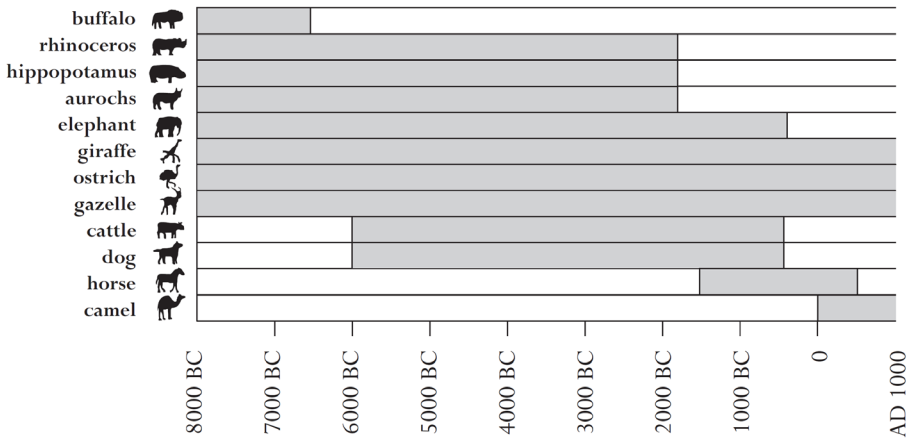


Figure B shows Saharan rock art (ancient drawings or paintings found on the walls of caves) and changing patterns of wildlife.

Figure B

Saharan rock art and changing patterns of wildlife



Source: Copyright Bartholomew Ltd 1988. Extracted from *The Times Atlas of Archaeology* and reproduced by permission of Harper Collins Publishers.

Use the information about Lake Chad on the opposite page to answer the questions below.

Question 1: LAKE CHAD (R040Q02)

***Reading task**: Retrieving information*
***Text format**: Non-continuous*
***Situation**: Public*

What is the depth of Lake Chad today?

(A) About two metres.

B About fifteen metres.

C About fifty metres.

D It has disappeared completely.

E The information is not provided.

Question 2: LAKE CHAD (R040Q03A)

***Reading task**: Retrieving information*
***Text format**: Non-continuous*
***Situation**: Public*

In about which year does the graph in Figure A start?

Question 3: LAKE CHAD (R040Q03B)

***Reading task**: Reflection and evaluation*
***Text format**: Non-continuous*
***Situation**: Public*

Why has the author chosen to start the graph at this point?

Question 4: LAKE CHAD (R040Q04)

***Reading task**: Interpreting texts*
***Text format**: Non-continuous*
***Situation**: Public*

Figure B is based on the assumption that

(A) the animals in the rock art were present in the area at the time they were drawn.

B the artists who drew the animals were highly skilled.

C the artists who drew the animals were able to travel widely.

D there was no attempt to domesticate the animals which were depicted in the rock art.

Question 5: LAKE CHAD *(R040Q06)*

Reading task: *Interpreting texts*
Text format: *Non-continuous*
Situation: *Public*

For this question you need to draw together information from Figure A and Figure B.

The disappearance of the rhinoceros, hippopotamus and aurochs from Saharan rock art happened

A at the beginning of the most recent Ice Age.

B in the middle of the period when Lake Chad was at its highest level.

Ⓒ after the level of Lake Chad had been falling for over a thousand years.

D at the beginning of an uninterrupted dry period.

Figure 4.1.1 Example of testlet assessing reading literacy

4.2 Longitudinal patterns of testlets usage across five waves of PISA

For the purpose of comparability of the cognitive scores across different waves, some items, and consequently testlets, were repeatedly used in different assessment waves and constituted the basis for linking (Gebhardt & Adams, 2007; Strietholt & Rosén, 2016) across time. Table 4.2.1 gives an account of various combinations of linking testlets and items used in different waves of PISA. The first five rows of the table report how many testlets (and corresponding questions) were used only once. The next four rows highlight testlets and items used in two PISA waves and so on. For example in 2012 when mathematics was targeted, the domain row labelled x_x_x_x_2012 reveals that 31 new testlets were introduced in this year with a total of 74 items. Similarly, when reading was the main cognitive domain for the second time in 2009, 14 new testlets comprised of a total of 53 items were introduced and not used again in 2012 (row x_x_x_2009_x in reading's column). The last row of the Table 4.2.1 (2000_2003_2006_2009_2012) accounts testlets that were employed in all of the five implementations of the PISA study. For example, there was only one reading testlet that fits this category which was reduced from five questions to three after PISA 2009. The table also offers a more complex overview. For example, looking at the last three rows approximately 27% (24 out of 89) of all mathematical testlets were reused for linking purposes, at least four times. At the same time, only about 13% (8 out of 63) of reading testlets were so frequently involved in the cross-wave linking. The table also indicates a preference for reusing the same testlets in subsequent waves as opposed to returning to them many years apart. For example, part of the table labelled "Used twice" suggests that only three reading testlets were being returned to after two assessments (nine years) time-out (pattern 2000_x_x_2009_x). The only other single item mathematics testlet that was not used in consecutive PISA implementations is listed in a row (x_2003_2006_x_2012) and was returned to in PISA 2012 after being dropped in PISA 2009.

Table 4.2.1 Distribution of testlets and items quantifying single assessment usage as well as ones used for linking

Number of times that the specific testlets were used and year of the PISA assessment in which there were used[15]		Mathematics		Reading		Science	
		Testlets	Items	Testlets	Items	Testlets	Items
USED ONCE	2000_x_x_x_x	**5**[16]	11_x_x_x_x	**26**	83_x_x_x_x	**4**	9_x_x_x_x
	x_2003_x_x_x	**16**	x_23_x_x_x	**0**		**1**	x_1_x_x_x
	x_x_2006_x_x	**0**		**0**		**14**	x_x_38_x_x
	x_x_x_2009_x	**0**		**14**	x_x_x_53_x	**0**	
	x_x_x_x_2012	**31**	x_x_x_x_74	**0**		**0**	
USED TWICE	2000_2003_x_x_x	**6**	12_12_x_x_x	**0**		**4**	11_11_x_x_x
	2000_x_x_2009_x	**0**		**3**	13_x_x_11_x	**0**	
	x_2003_2006_x_x	**6**	x_12_12_x_x	**0**		**1**	x_4_4_x_x
	x_x_x_2009_2012	**0**		**12**	x_x_x_41_41	**0**	
USED THREE TIMES	2000_2003_2006_x_x	**0**		**0**		**3**	8_8_8_x_x
	x_2003_2006_2009_x	**0**		**0**		**0**	
	x_2003_2006_x_2012	**1**	x_1_1_x_1	**0**		**0**	
	x_x_2006_2009_2012	**0**		**0**		**14**	x_x_43_43_43
USED FOUR TIMES	2000_2003_2006_2009_x	**0**		**7**	28_23_23_21_x	**0**	
	x_2003_2006_2009_2012	**19**[17]	x_28_27_27_26	**0**		**1**	x_4_4_4_4
USED FIVE TIMES	2000_2003_2006_2009_2012	**5**	8_8_8_8_8	**1**	5_5_5_5_3	**3**	6_6_6_6_6

[15] Symbol "x" indicates that specific assessment year is not under consideration.
[16] For example, this number shows that in PISA 2000 there were 5 mathematical testlets (constituting 11 items) which were never used again in the future studies.
[17] For example, this number pinpoints that 19 mathematical testlets were used in last four waves of PISA study. Furthermore, some of the testlets were adjusted with time by dropping one question in PISA 2006 and another one in PISA 2012.

4.3 Within-testlet variability of item difficulty estimates

In order to show a graphical representation of within-testlet variability Figure 4.3.1, 4.3.2 and 4.3.3 (for mathematics, reading and science, respectively)[18] show the range of the percentage of correct responses in international samples of students for cognitive items within testlets. The horizontal axis lists all testlet names, with various colours denoting the number of items in each testlet. A minus sign[19] indicates the average percentage of correct responses within each testlet with lower and upper limits representing the within-testlet maximum and minimum difficulties, respectively. Testlets are sorted left to right by the average difficulty from the easiest to the most difficult.

[18] To facilitate the in depth review of these Figures, the electronic pdf versions are offered with items also listed alphabetically. (Figure 4.3.1, Figure 4.3.1_Alphabetic order, Figure 4.3.2, Figure 4.3.2_Alphabetic order, Figure 4.3.3, Figure 4.3.3_Alphabetic order).

[19] Each time when only a single "-" is reported this indicates that was only one item labelled by a specific testlet name.

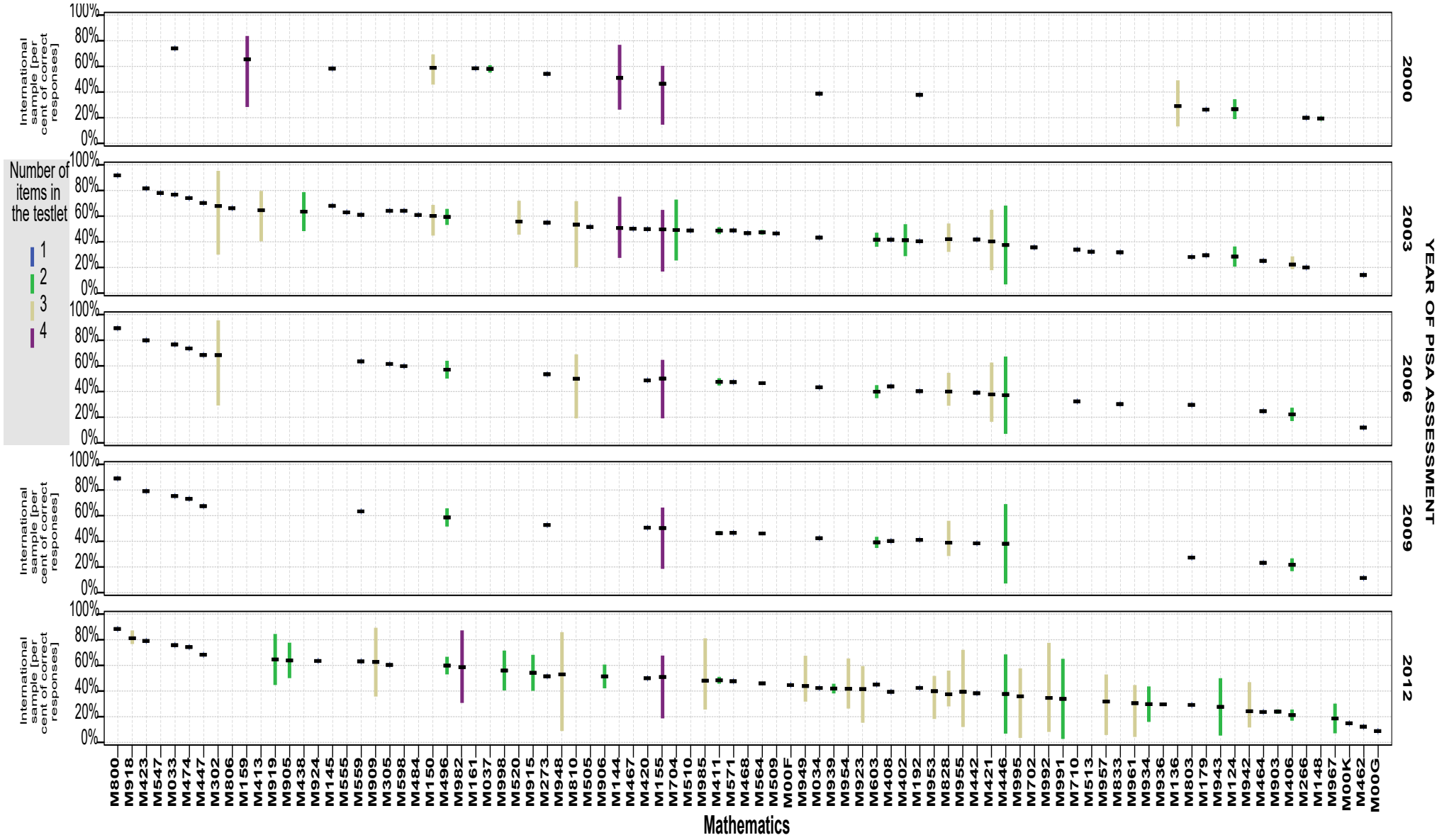


Figure 4.3.1 Range of the percentage of correct responses for mathematics items within each testlet and across five PISA assessments

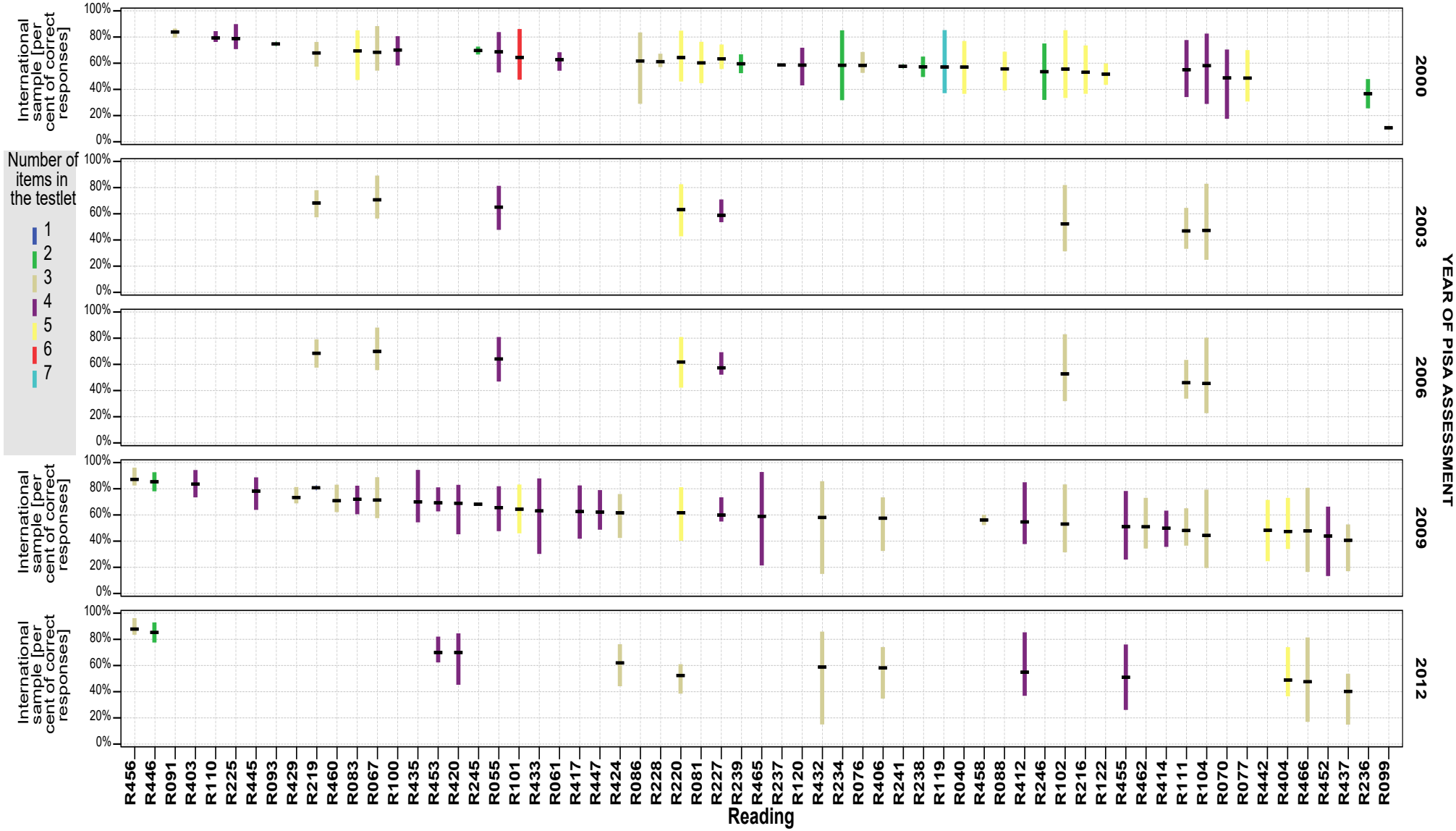


Figure 4.3.2 Range of the percentage of correct responses for reading items within each testlet and across five PISA assessments

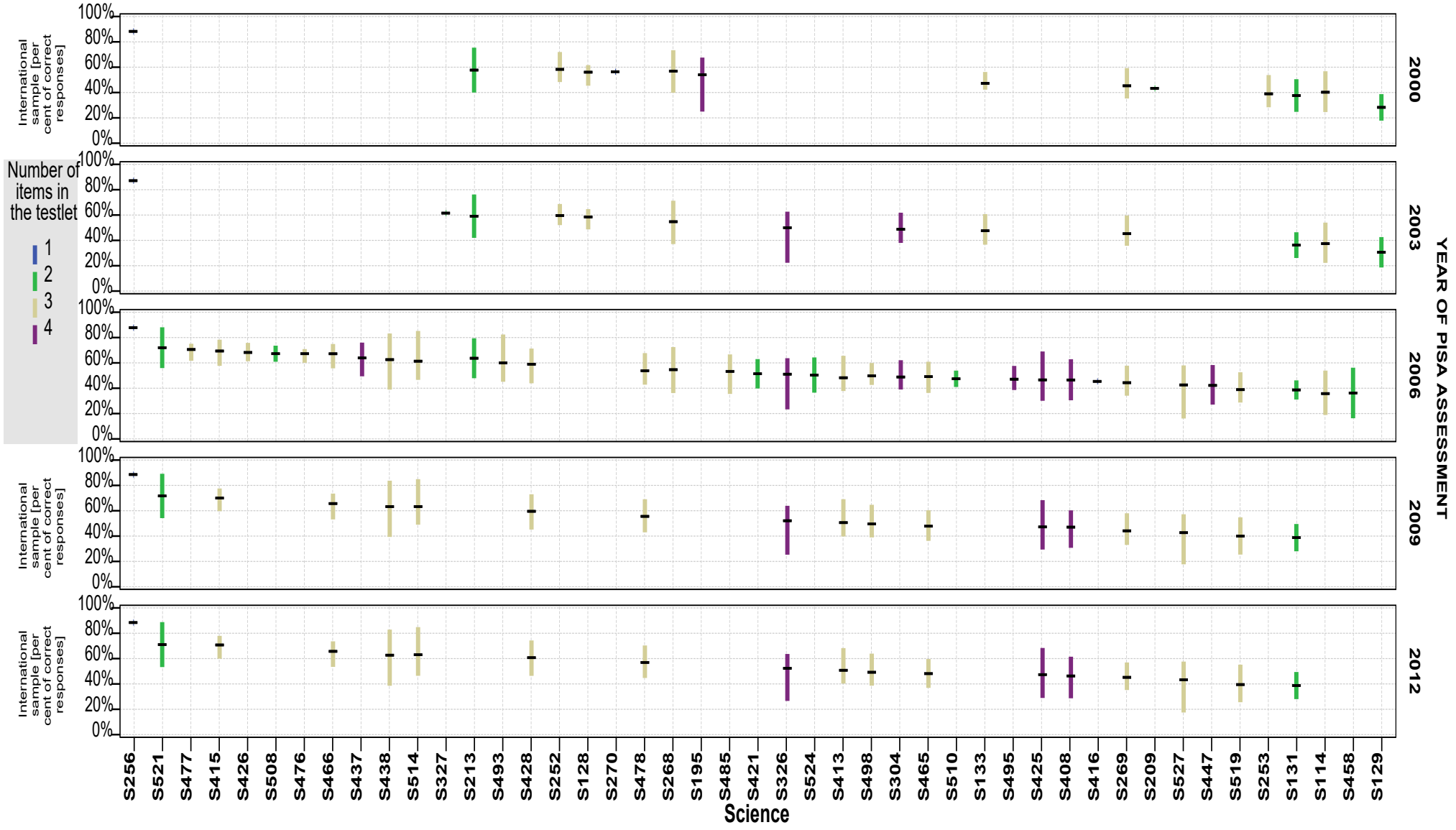


Figure 4.3.3 Range of the percentage of correct responses for science items within each testlet and across five PISA assessments

In addition to showing the range of item difficulties within testlets, all three figures provide a cross-wave overview of the testlets used in linking different PISA waves. For example, Figure 4.3.1 includes the mathematical four-item testlet M155 "Population Pyramids" which was employed in all five PISA studies investigated in this research project. Single item testlets M033 "A View with a Room", M034 "Bricks", M192 "Containers" and M273 "Pipelines" also featured in all five PISA implementations. A number of other testlets were used four times, such as the two-item testlets M446 "Thermometer Cricket" and M406 "Running Tracks"[20]. Interestingly, these two testlets are quite different in item composition. Testlet M446 consists of a pair of items with one being relatively easy (about 70% of correct responses consistent across time) and the second one being quite difficult. By contrast, both items from M406 were challenging for the international student samples. From Figure 4.3.1, it can be seen that in 2003 and 2012 mathematics was the targeted cognitive domain as the number of items used is larger compared to other three PISA waves.

Figure 4.3.2 presents reading testlets and their use for the purpose of cross-wave linking. Reading testlets had more items per testlet compared to mathematics. For example, R119 "Employment" and R101 "Rhinoceros" comprised seven and six questions, respectively. The number of items per testlet was reduced after the first wave, and the size of linking testlets was reduced in subsequent PISA waves as represented by R102 "Shirts" and R104 "Telephone". Cross-wave linking in the second and third PISA waves was based on the same eight testlets of which only R220 "South Pole" was used in 2012. This reading testlet is the only one that was presented to students in all five PISA waves investigated in this study.

The distribution of testlets from science is shown in Figure 4.3.3. Three testlets: S131 "Good Vibrations", S269 "Earth's Temperature" and S256 "Spoons" were used in PISA 2000, 2003, 2006, 2009 and 2012. The maximum number of items in science testlets was four, and that was similar to mathematics. However, the majority of science testlets were three or four-item large while only a few mathematical testlets has four questions. The same set of science testlets was used in PISA 2009 and 2012.

A review of all three figures indicates that single item testlets were used extensively only in testing mathematical literacy. Reading testlets had on average the greatest number of items per testlet with a large proportion being four-item or larger. Only a small number of reading

[20] Testlet M406 started in PISA 2003 with three items, but one of them was abandoned after this wave.

and science questions was correctly answered by fewer than 20% of students from the international samples, while the same cannot be said for mathematical items. Testlets used for cross-wave linking appear to have relatively constant variation across time. In science, the same sized testlets have comparable difficulty ranges. However, in mathematics or reading, the same size testlets can be considerably different by occasionally including questions of a similar difficulty or on another occasion showing large difficulty ranges. For example in PISA 2009 testlets R453 “Find Summer Job” and R465 “Different Climates” show this characteristic.

4.4 Summary

This chapter gives descriptive information about the number of items that were used in all waves with particular focus being placed on testlets of items under common introduction or passage. The results reported in this chapter facilitate a graphical overview of testlets used in different implementations of PISA studies. Such an overview is essential because despite extensive technical information about each separate PISA study (Adams & Wu, 2002; OECD, 2005b, 2009b, 2012, 2014b) there are no publications which give researchers a detailed overview of the items re-used for PISA cross-wave linking. Worth highlighting from this chapter is the limited number of testlets (only eight) being used in reading assessment in PISA 2003 and 2006, serving the purpose of linking four PISA waves. The differences in the size of the testlets are also worth noting with a large proportion of single item testlets used in mathematics while larger testlets, of up to seven items, were used in reading. This chapter also offered insights into the within-testlet spread of item difficulties with some of the same sized testlets being very homogeneous while other testlets including items of considerably varying difficulty.

CHAPTER 5 RESULTS FOR RESEARCH AIM 2 - LID IN THE PISA INTERNATIONAL CALIBRATIONS

5.1 The organisation of the chapter

This chapter addresses research questions involving the data from international calibrations composed of cognitive datasets from 26 OECD countries for the five waves from 2000 to 2012 inclusive and for the three cognitive domains: reading, mathematics and science literacy. The chapter begins with a section 5.2 offering the estimates of the overall prevalence of LID and considers its prevalence across the three cognitive domains. This is followed by a section 5.3 reporting LID presence by PISA wave, cognitive domain, and the location (within or between testlets) of item pairs. Section 5.4 consists of five key subsections that attempt to explain the plausible causes of LID. Three sub-sections in 5.4.1, which are dedicated to mathematics, reading and science, are descriptive in nature utilising the items and testlets that have been released publicly. The qualitative investigation also serves the purpose of finding which testlets show a cross-wave consistency in the presence of the local item dependency. These qualitative overviews are followed by two key sections in 5.4.2 that also offer suggestions for drivers of item dependency. However, in these sections, the investigations are based on statistical analyses that include information about the items including, for example, their levels of difficulty, source, language submitted, and question format. In addition, item characteristics driven by the PISA assessment frameworks are included in these analyses. Through these analyses, the influence of these item characteristics on LID is investigated. The pdf bookmarks are provided to facilitate targeted navigation of the chapter.

5.2 Is LID present in any of the international calibrations data?

This section offers an overview of the existence of LID in PISA by reporting the prevalence of LID according to cut points proposed in the methodology chapter (see section 3.3.3.2). The discussion starts with reporting the fit indices for all Confirmatory Factor Analyses (CFAs) that were undertaken with the international calibrations samples. This is followed by a graphical overview of dependency existence in the form of box-plots produced for each cognitive domain and PISA wave. The section concludes by reporting the percentage of item pairs that indicate a violation of the assumption of local item independence. Table 5.2.1 reports overall fit statistics for all CCFAs from three domains and all five waves of PISA.

Table 5.2.1 Fit statistics from Confirmatory Factor Analyses undertaken with international calibration data for three cognitive domains and five PISA waves.

	PISA 2000	**PISA 2000**	**PISA 2003**	**PISA 2006**	**PISA 2009**	**PISA 2012**	**PISA 2000**	**PISA 2000**	**PISA 2000**	**PISA 2003**	**PISA 2006**	**PISA 2009**	**PISA 2012**	**PISA 2000**	**PISA 2000**	**PISA 2003**	**PISA 2006**	**PISA 2009**	**PISA 2012**
	M_O1[21]	M_O2	M	M	M	M	R_O1	R_O2	R_O3	R	R	R	R	S_O1	S_O2	S	S	S	S
Number of Free Parameters	56	56	176	100	73	176	211	111	68	62	62	209	89	54	53	70	212	109	109
Chi-Square Test of Model Fit	865	1643	7150	2462	1668	5870	8959	3306	1546	1992	1652	9968	2849	601	907	1521	7305	2673	2363
Degrees of Freedom	230	230	3402	1080	560	3402	4559	1274	464	350	350	4949	902	275	275	527	5150	1325	1325
P-Value	p<0.001	p<0.001	p<0.001	p<0.001	p<0.001	p<0.001	p<0.001	p<0.001	p<0.001	p<0.001	p<0.001	p<0.001	p<0.001	p<0.001	p<0.001	p<0.001	p<0.001	p<0.001	p<0.001
Chi-Square / Degree of Freedom	3.8	7.1	2.1	2.3	3.0	1.7	2.0	2.6	3.3	5.7	4.7	2.0	3.2	2.2	3.3	2.9	1.4	2.0	1.8
RMSEA	0.02	0.03	0.01	0.01	0.02	0.01	0.01	0.01	0.02	0.03	0.02	0.01	0.02	0.01	0.02	0.02	0.01	0.01	0.01
90 Percent C.I.	0.018 0.021	0.028 0.031	0.009 0.010	0.011 0.012	0.014 0.016	0.007 0.008	0.009 0.010	0.011 0.012	0.022 0.025	0.025 0.027	0.023 0.025	0.009 0.009	0.015 0.016	0.012 0.014	0.017 0.019	0.015 0.017	0.005 0.006	0.010 0.011	0.009 0.010
Probability RMSEA <= .05	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
CFI	0.98	0.96	0.98	0.98	0.97	0.98	0.97	0.96	0.98	0.95	0.97	0.97	0.96	0.99	0.98	0.98	0.99	0.98	0.98
TLI	0.98	0.95	0.97	0.98	0.97	0.98	0.97	0.96	0.97	0.95	0.97	0.97	0.96	0.99	0.97	0.98	0.99	0.98	0.98
WRMR	1.67	2.36	1.53	1.47	1.63	1.35	1.47	1.57	1.51	2.20	2.00	1.56	1.79	1.28	1.60	1.56	1.21	1.38	1.29

[21] Reasons for two or three estimations in PISA 2000 labelled as O1, O2 or O3 offered in Methodology (see Section 3.3.2.3)

The fit of the models presented in Table 5.2.1 is acceptable for all analyses according to RMSEA, CFI, and TLI (Schreiber, Stage, King, Nora, & Barlow, 2006) while acknowledging literature regarding their suitability for categorical variables (Garrido, Abad, & Ponsoda, 2016; Heene, Hilbert, Freudenthaler, & Bühner, 2012; Maydeu-Olivares & Joe, 2014; Nye & Drasgow, 2011). Interestingly, if the WRMR threshold rule of <1 proposed by Yu (2002) was adhered to, all the models would be misfitting. Cook, Kallen, and Amtmann (2009) suggest that WRMR may be more sensitive to multidimensionality. The statistically significant Chi-square values could also be used to argue against the reported univariate CFAs. This statistic is very frequently ignored in practice by pointing to its over-sensitivity for large models utilising large sample sizes. At the same time, the specialised structural equation modelling forum SEMNET recorded extensive discussion as to whether this oversimplification should be exercised, with both sides of the argument failing to convince each other. Finally, should a fit rule of the ratio of chi-square to degrees of freedom not exceeding three be applied (Schreiber et al., 2006), mixed conclusions would arise with over 40% of CFAs being questionable according to this canon.

Table 5.2.2 reports the proportions of all pairs of items for which residual correlations exceed the absolute value of 0.1. This table quantifies the LID prevalence across all combinations of PISA waves and domains elaborating on LID-indicated pairs of items conditional on the sign of RCs and whether the domain tested was targeted in the specific PISA wave.

Table 5.2.2 Proportions of all pairs of items for which residual correlations either exceeds the absolute value of 0.1, are less than -0.1 or are more than +0.1

		Absolute value of RC exceeds 0.1		RC exceeds +0.1		RC is less than -0.1		Total number of RCs[i]	Total number of students used in CFAs[ii]
		Count	Row%	Count	Row%	Count	Row%		
Mathematics	2000	60	**15.0%**	18	**4.5%**	42	**10.5%**	401	7080 (O1) / 7157 (O2) [iii]
	2003	406	**11.6%**	121	**3.5%**	285	**8.2%**	3486	12893
	2006	117	**10.4%**	45	**4.0%**	72	**6.4%**	1128	9918
	2009	75	**12.6%**	14	**2.4%**	61	**10.3%**	595	9084
	2012	393	**11.3%**	142	**4.1%**	251	**7.2%**	3486	12907
Science	2000	31	**6.5%**	8	**1.7%**	23	**4.8%**	480	10105 (O1) / 12933 (O2) / 4305 (O3) [iii]
	2003	43	**7.7%**	9	**1.6%**	34	**6.1%**	561	6992
	2006	358	**6.8%**	125	**2.4%**	233	**4.4%**	5253	12987
	2009	114	**8.3%**	35	**2.5%**	79	**5.7%**	1378	9071
	2012	107	**7.8%**	26	**1.9%**	81	**5.9%**	1378	8956
Reading	2000	370	**6.6%**	126	**2.2%**	244	**4.3%**	5632	7027 (O1) / 7139 (O2) [iii]
	2003	40	**10.6%**	10	**2.6%**	30	**7.9%**	378	6942
	2006	28	**7.4%**	13	**3.4%**	15	**4.0%**	378	6610
	2009	821	**16.3%**	203	**4.0%**	618	**12.2%**	5050	12988
	2012	195	**20.6%**	37	**3.9%**	158	**16.7%**	946	8870

[i] In some rows counts do not match number against a formula (number of items)*(number of items-1)/2 if the number of items data from Table 4.1.1 was to be used. For all PISA 2000 studies, the unbalanced booklet design explains the discrepancy. For reading 2009 and mathematics 2012 use of the easy booklet option caused selected 26 OECD countries do not respond to all items.

[ii] While 500 students were randomly sampled from 26 OECD countries the students' sample sizes are smaller than 13000. There are three reasons for this. (1) For non-targeted domains, some students were not exposed to cognitive questions from all three cognitive domains. (2) The simple random subsamples of 500 students from each country's datasets were stratified by different strata variables following the procedures indicated in the PISA Technical Manuals (OECD, 2014b, p.163). The rounding involved with stratification resulted in country level samples not always being equal 500. (3) In PISA 2006 due to an error in printing the booklets in the USA, the PISA consortium decided to exclude the American reading data from the cognitive database (OECD, 2007a)

[iii] Reasons for two or three numbers in PISA 2000 labelled as O1, O2 or O3 offered in Methodology (see Section 3.3.2.2)

To offer combined estimates, a meta-analytic approach was undertaken. The section below reports event rates and their 95% confidence intervals. Comprehensive Meta-Analysis software was used (Borenstein et al., 2014) with a random effects model implemented to combine results within each cognitive domain. The presence of LID is less pronounced for science with a cross wave meta-analytical rate of 7.3% (6.7%, 8.0%) compared to mathematics with across wave rates of 11.7% (10.8%, 12.7%). Corresponding result for reading is 11.4% (6.8%, 18.4%). While focusing on each domain cross wave prevalence for high positive RCs are 3.8% (3.3%, 4.3%), for mathematics, 2.3% (2%, 2.6%) for science and 3.2% (2.3%, 4.4%) for reading. The equivalent negative RCs estimates are 8.2% (7%, 9.5%) for mathematics, 5.3% (4.5%, 6.1%) for science and 8% (4.4%, 13.9%) for reading. While in many publications, assessment of reading literacy is expected to return LID due to the use of common reading passages, the results presented in Table 5.2.2 indicate that LID presence is just as frequent in mathematics as it is for reading. Worth noting is the fact that outlying large negative residual correlations are, on average across all 15 rows, 2.5 times more prevalent than positive outlying RCs. However, this ratio is very close to one in the case of reading in 2006 and exceeding four for mathematics in 2009 and reading in 2012.

It is of interest to elaborate on the impact of the greater number of items for the targeted domains in each PISA wave. The underline in the table highlights the year in which a specific cognitive domain was targeted. It can be seen that the number of items used in PISA assessments appears to be unrelated to LID prevalence. In PISA 2000 and 2009 over 100 items were used in the estimation of reading RCs, and there is close to 10% discrepancy between these two PISA implementations in the proportion of RCs exceeding an absolute value of 0.1. Furthermore, the largest percent (20.6%) of outlying RCs is also reported for reading in 2012 when this cognitive domain was not targeted in this PISA wave and involved 44 items.

In conclusion, the presence of LID is observed, and investigation of its possible causes is warranted. LID prevalence does vary by cognitive domain with results for mathematics being similar to reading yet lower for science. The percentage of outlying positive residual correlations is consistently lower compared to negative RCs.

It is important to highlight the approach undertaken in this study of treating not-reached cognitive questions, i.e. “all consecutive missing values starting from the end of each cognitive session” (Adams & Wu, 2002, p. 130) was to code them as missing rather than

coding them as failed. The method used in this study for dealing with not-reached items is likely to be conservative in regard to dependency prevalence. For example, Monseur et al. (2011, p. 139) found that for some PISA 2000 reading testlets, coding not-reached as "failed", produced average residual correlations twice the size compared with the case in which not-reached are treated as missing. The method of dealing with not-reached items changed after PISA 2012, treating them exclusively as not administered. Before PISA 2015 a dual approach was applied in which not-reached responses "were considered as wrong answers when estimating student proficiency (i.e. in the "scoring" step) but as not administered when estimating item parameters (in the "scaling" step)" (OECD, 2016a, p. 306).

5.3 Does the prevalence of LID vary by item pair location?

This section expands on the information presented above. The previous section addressed overall LID prevalence. In this section, the location of item pairs, either both within a testlet or in different testlets, is investigated.

5.3.1 The prevalence of positive and negative LID for within-testlet pairs of cognitive items.

Table 5.3.1 shows numbers and proportions of item pairs with positive and negative RCs for within-testlet item pairs by PISA wave and domain. The more frequent utilisation of singular testlets in mathematics, presented above in Figure 4.3.1, is also confirmed here by lower numbers of mathematics within-testlet pairings.

Table 5.3.1 Percentage of within-testlet pairs indicating positive and negative LID, by cognitive domain and PISA wave.

		RC exceeds 0.1		RC is lower than -0.1		Total number of within-testlet pairs of items
		Count	Row N %	Count	Row N %	
Mathematics	2000	12	46%	0	0%	26
	2003	30	67%	0	0%	45
	2006	7	29%	0	0%	24
	2009	6	40%	0	0%	15
	2012	16	30%	1	2%	53
Science	2000	2	7%	0	0%	30
	2003	5	15%	0	0%	33
	2006	11	10%	0	0%	107
	2009	4	7%	0	0%	56
	2012	2	4%	0	0%	56
Reading	2000	49	25%	0	0%	199
	2003	10	27%	0	0%	37
	2006	11	30%	0	0%	37
	2009	39	28%	0	0%	138
	2012	13	23%	0	0%	56

The Table 5.3.1 suggests that despite fewer non-singular testlets being used in mathematics, ones which were given to students offered larger proportions of positive within-testlet LID as compared to reading and science. The same meta-analytical approach reported in section 5.2 was also implemented in this section revealing positive within-testlet LID prevalence in the mathematics of 43% (28%, 59%). The same prevalence showed a lower tendency for reading, being equal to 26% (22%, 30%) and much smaller for science with 9% (6%, 13%). Only one pair of items (M998Q02 and M998Q04T) from the same testlet reported RC lower than -0.1, and it came from the non-released mathematical testlet M998 called “Bike rental”.

Looking at the same topic from testlet level data, out of the 39 non-singular mathematics testlets used throughout the five waves of PISA[22], 27 (69%) had at least one pair of its items for which the RC exceeded +0.1. A similar proportion of 71% (37 out of 52 non-singular testlets) was found to have at least one item pair with a positive RC for reading. In science, 12 non-singular testlets out of 40 (30%) had at least one pair of dependency indicating items. Thus, for reading and mathematics, we find that a majority of multi-item testlets include pairs

[22] Based on international calibration samples from 26 OECD countries

of items that violate the LII assumption. For science, the incidence of this violation is smaller.

5.3.2 The prevalence of positive and negative LID for between-testlet pairs of cognitive items

The meta-analytic prevalence of positive LID for between-testlet pairs is shown in Table 5.3.2. For mathematical questions this prevalence was 2.8% (2.1%, 3.6%), while for reading and science items it was 1.8% (1%, 3.2%) and 2% (1.5%, 2.5%), respectively. It is worth noting that the science estimate for between-testlet LID does not differ greatly from other cognitive domains in contrast to results in the previous section, 5.3.1, for within-testlet LID. Percentages of RCs lower than -0.1 out of all between-testlet item pairs varied more for reading resulting in larger meta-analytical confidence intervals. The prevalence of negative LID for between-testlet items couplings was 8.4% (7.1%, 9.8%) for mathematics, 8.5% (4.7%, 14.8%) for reading and 5.5% (4.7%, 6.5%) for science.

Table 5.3.2 Percent of between-testlet pairs indicating positive and negative LID, by cognitive domain and PISA wave.

		RC exceeds 0.1		RC is lower than -0.1		Total number of between-testlet pairs of items
		Count	Row N %	Count	Row N %	
Mathematics	2000	6	1.6%	42	11.2%	375
	2003	91	2.6%	285	8.3%	3441
	2006	38	3.4%	72	6.5%	1104
	2009	8	1.4%	61	10.5%	580
	2012	126	3.7%	250	7.3%	3432
Science	2000	6	1.3%	23	5.1%	450
	2003	4	0.8%	34	6.4%	528
	2006	114	2.2%	233	4.5%	5146
	2009	31	2.3%	79	6.0%	1322
	2012	24	1.8%	81	6.1%	1322
Reading	2000	77	1.4%	244	4.5%	5433
	2003	0	0.0%	30	8.8%	341
	2006	2	0.6%	15	4.4%	341
	2009	164	3.3%	618	12.6%	4912
	2012	24	2.7%	158	17.8%	890

5.3.3 The prevalence of positive LID among residual correlations for which absolute value exceeds 0.1

This section focuses on the denominator of all RCs exceeding an absolute value of 0.1. All PISA waves' results, which can be viewed in Table 5.3.3, are aggregated using a meta-analytical approach. The prevalence of RCs indicating positive LID is reported in Table 5.3.4

for item pairs within testlets and between testlets

Table 5.3.3 Proportions of all pairs of items for which residual correlations are either higher than +0.1 or lower than -0.1, by item pairs testlet placement domain and wave

			RC is lower than -0.1				RC exceeds 0.1			
			Count	Row N %	95.0% Lower CL for Row N %	95.0% Upper CL for Row N %	Count	Row N %	95.0% Lower CL for Row N %	95.0% Upper CL for Row N %
Between-testlet placement	**Maths**	2000	42	88%	76%	95%	6	12%	5%	24%
		2003	285	76%	71%	80%	91	24%	20%	29%
		2006	72	66%	56%	74%	38	34%	26%	44%
		2009	61	88%	79%	94%	8	12%	6%	21%
		2012	250	67%	62%	71%	126	33%	29%	38%
	Science	2000	23	79%	62%	91%	6	21%	9%	38%
		2003	34	90%	77%	96%	4	10%	4%	23%
		2006	233	67%	62%	72%	114	33%	28%	38%
		2009	79	72%	63%	80%	31	28%	20%	37%
		2012	81	77%	68%	84%	24	23%	16%	32%
	Reading	2000	244	76%	71%	80%	77	24%	20%	29%
		2003	30	100%			0	0%		
		2006	15	88%	67%	98%	2	12%	3%	33%
		2009	618	79%	76%	82%	164	21%	18%	24%
		2012	158	87%	81%	91%	24	13%	9%	19%
Within-testlet placement	**Maths**	2000	0	0%			12	100%		
		2003	0	0%			30	100%		
		2006	0	0%			7	100%		
		2009	0	0%			6	100%		
		2012	1	6%	1%	24%	16	94%	76%	99%
	Science	2000	0	0%			2	100%		
		2003	0	0%			5	100%		
		2006	0	0%			11	100%		
		2009	0	0%			4	100%		
		2012	0	0%			2	100%		
	Reading	2000	0	0%			49	100%		
		2003	0	0%			10	100%		
		2006	0	0%			11	100%		
		2009	0	0%			39	100%		
		2012	0	0%			13	100%		

Table 5.3.4 Meta-analytic prevalence of residual correlations higher than +0.1 taken out of all LID indicative item pairs, by testlet placement and cognitive domain

		Number of RCs lower than -0.1	Number of RCs exceeding 0.1	Meta-analytical prevalence of positive RCs	Meta-analytical Confidence Interval of prevalence of positive RCs	Total RCs exceeding absolute RCs value of 0.1
Between-testlet placement	Mathematics	710	269	**24%**	(17%,33%)	979
	Science	450	179	**25%**	(19%,33%)	629
	Reading	1065	267	**19%**	(14%,24%)	1332
Within-testlet placement	Mathematics	1	71	**95%**	(87%,99%)	72
	Science	0	24	**90%**	(71%,97%)	24
	Reading	0	122	**98%**	(92%.99%)	122

The total number of item pair RCs for which the absolute value exceeded 0.1 was 3158, and 218 of these originated from within-testlet pairs of items.

Of the 218 within-testlet RCs, 217 were positive, i.e. indicative of positive LID. A considerable proportion, 715 out of the 2940 item pairs had positive RCs that came from pairs of items from different testlets. For reading, the corresponding prevalence estimate is somewhat lower 19% (14%,24%) as compared to mathematics and science with 24% (17%,33%) and 25% (19%,33%), respectively.

In answer to the research question posed in this section, 'Does the prevalence of LID vary by cognitive domain, PISA wave or item pair location?' the conclusion is affirmative on all counts. Thus, subsequent analyses that aim to offer plausible explanations for the LID causes have to control for the factors of wave, item pair location, and cognitive domain. Consequently, Section 5.4 attempts to offer in-depth explanations for the presence of LID.

5.4 Possible causes for LID?

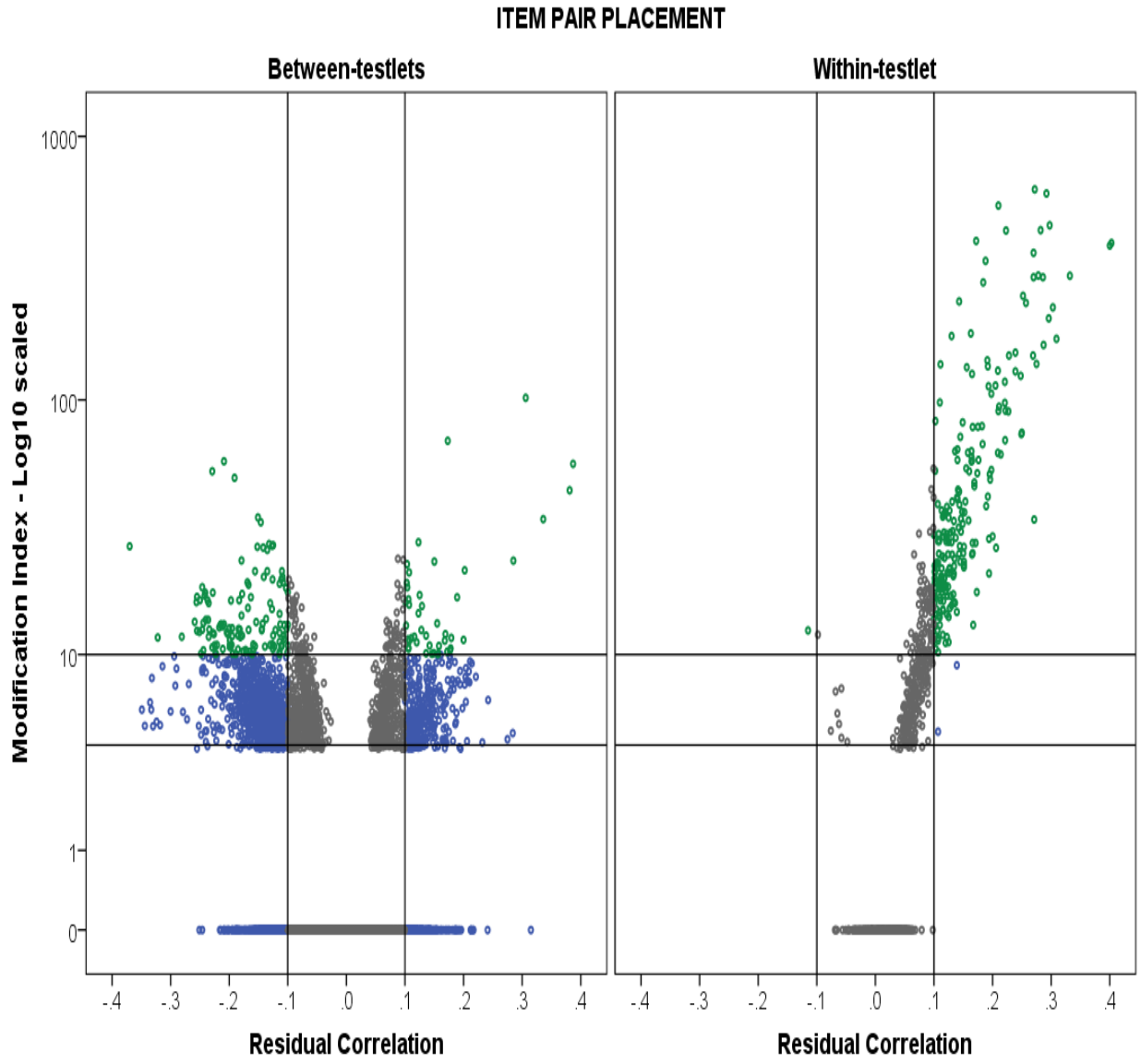
The first half of this section takes advantage of some items being released, permitting a descriptive overview of LID drivers. The examination identifies whether the LID is positive or negative and whether the item pairs are located within a testlet or between testlets. Furthermore, characteristics of the items, e.g. the sub-domain that the items test, the item

format, or the particular skill that is tested by the items, are evaluated. The evaluation is qualitative and is undertaken in order to generate possible explanations for the LID that is observed. The second half offers a quantitative picture taking advantage of all available information about the non-released items, using multilevel logistic regression to determine what factors appear to be more important in explaining LID presence.

5.4.1 Qualitative investigation of LID drivers based on released PISA items along with and an overview of a cross-wave LID consistency

Section 5.2 reports the substantial prevalence of RCs exceeding an absolute value of 0.1. The aim of this part of Chapter 5 is to meaningfully graphically present pairs of items indicating LID and reproduce for the reader items which are released to the public, focusing on generating plausible explanations as to what may be causing the LID. In order to limit the number of pairs of items plotted and reproduced in this section another non-IRT based LID index, namely the modification index, was also incorporated in addition to Kline's (2016) RCs rule of thumb. While various modification index (MI) cut-points are utilised in the literature, this research follows the Mplus default value of 10[23]. Figure 5.4.1 indicates that almost all within-testlet positive large RCs (>=0.1) also returned MIs exceeding 10. For the sake of differentiating the pairs of items investigated with two LID indictors, the term dual-index LID will be used throughout this section.

[23] The minimum of two modification indices was used for a pair of items, which in PISA 2000 were estimated twice.



NOTE: Green colour represents dual-index LID, while blue shows residual correlations with an absolute value exceeding 0.1 and modification indices less than 10. Lack of graph continuity for MIs less than 4 is an artefact of not requesting them to be estimated and later recording them as MI=0 for the sake of being reported in this figure.

Figure 5.4.1 Scatterplot of residual correlations against modification indices, by item pair placement

The remainder of this section focuses on those item pairs displaying dual-index LID (green data points in Figure 5.4.1). These are the 385 item pairs for which the absolute values of RCs are greater than 0.1 and MIs are grater than 10. The distributions of high positive and negative RCs are investigated within and between testlets. A total of 15 graphs visualising 385 RCs are reported below. The graphs were produced using the qgraph package for R (Epskamp et al., 2012). This package allows the visualisation of data by applying network modelling techniques that, in this research, were applied to the matrix of RCs for all cognitive domains and waves. The graphs include lines for item pairs with RCs >= 0.1 and MIs >=10

in green and RCs <= -0.1 and MIs>=10 in red. Items from the same testlets are shown in the same shade. The figures are produced in high graphical resolutions allowing to zoom on desired sections facilitating their legibility and making the relationships between items much more obvious. This is particularly helpful for the target cognitive domains in each wave, e.g. mathematics in PISA 2003. All graphs generated with qgraph are accessible via links embedded in the test of this chapter. Furthermore, the electronic appendices also offer network plots for item testlets in which original LID, based only on RCs, is identified. (This information is reported in Section 5.2). These extended figures are reported for the sake of consistency with the previous section, but they have limited usability in the main targeted domains due to a large number of lines plotted. In these network plots, items from the same testlet are shown in the same shade, the strength of the RCs in reflected in the thickness of the lines, and the sense of the RC is colour coded, green for positive and red for negative RCs. The item pairs showing dual-index LID are also featured in corresponding MS Excel files which reproduce all released items. This sub-section is structured by cognitive domain and PISA wave (2000 to 2012) and by a combination of LID sign (positive or negative) and item pair location (within- or between-testlets). This hierarchical approach to reporting aims to facilitate quick and selective access to a specific PISA year or domain for the reader who also may be particularly interested in specific instances of the PISA study. Summaries concluding each cognitive domain sub section focus on relating the results of this study to existing literature reporting LID in PISA. The summaries also concentrate on the discussion about cross-wave consistency in the presence of the local item dependency.

5.4.1.1 Qualitative investigation of reasons for LID in the mathematics domain

The qualitative reporting on the presence of LID begins with the mathematics domain. This choice is driven by the large proportion of items that were released for mathematics. The format of reporting LID is the same for all three domains by showing LID pairs of items organised by a wave with a subdivision focusing on positive LID within the testlet, positive LID between the testlets and negative LID between the testlets. The text is supplemented by electronic resources.

PISA 2000

Figure 5.4.2, which is also available in an electronic version (Electronic Figure 5.4.2) reports LID occurrence between mathematical items in the PISA 2000 wave.

Green lines – positive residual correlations
Red lines –negative residual correlations

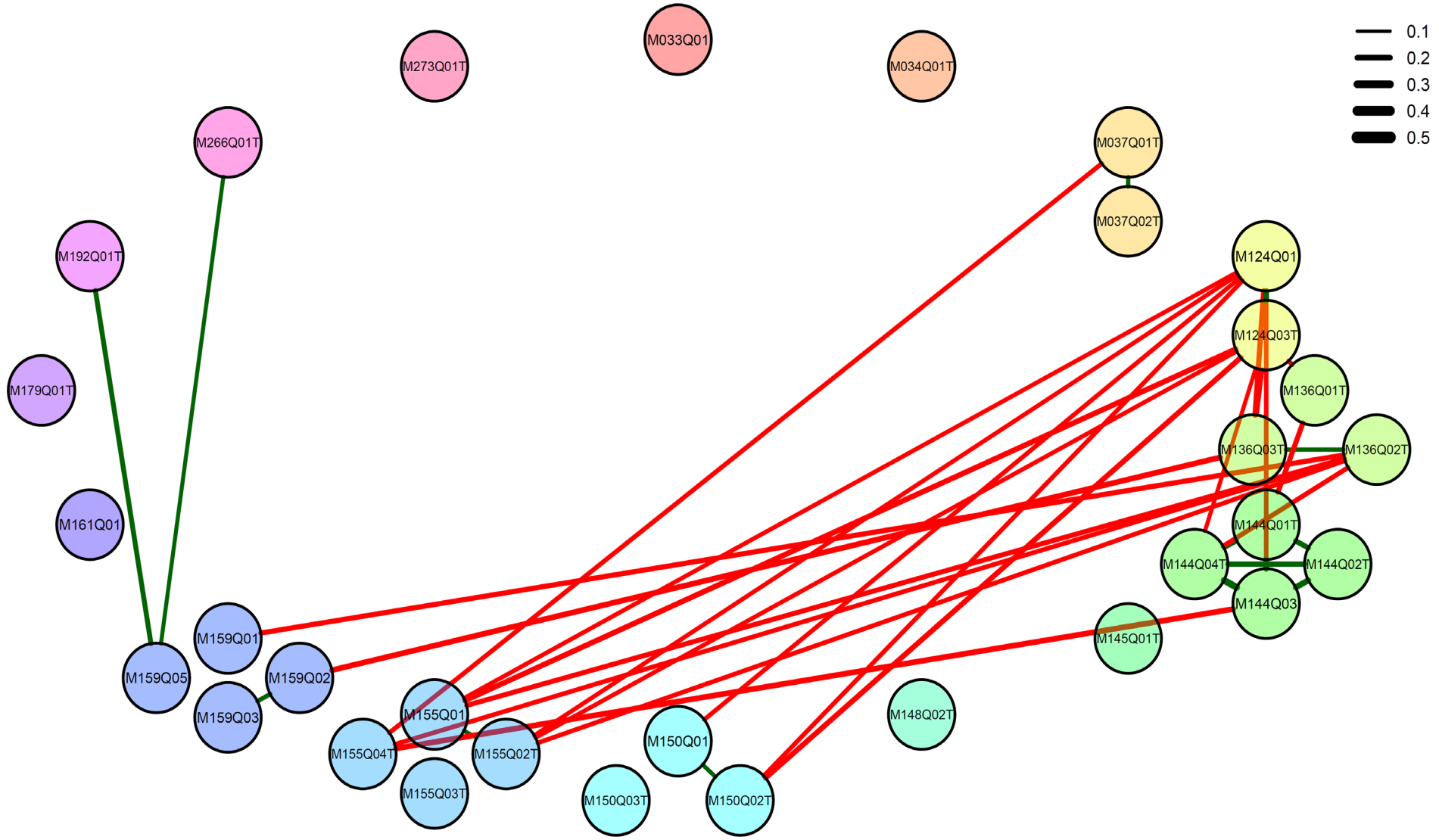


Figure 5.4.2 Visualisation of residual correlations data as a network - Mathematics PISA 2000

Positive LID between pairs of items within the same testlets

All non-singular testlets have at least one pair of items indicating positive LID between their items. Of particular interest is the testlet "Cube Painting" coded in PISA as M144 as each combination of its item pairs indicates positive LID. This testlet has not been released, so it is not possible to assert particular causes for the observed LID. A possible explanation for the LID is the use of a common stimulus. This cannot, however, be exclusively hypothesised as all four M144 items come from the same mathematical strand (Geometry), mathematical concept (Space and shape) and item context (Educational) as per the PISA 2000 assessment framework. Interestingly, five out of six remaining pairs of items with high positive LID come from testlets that were released to the public. This gives an opportunity of investigating whether LID is driven largely by common stimuli or some other cause. Electronic appendix (Electronic Appendix for Figure 5.4.2_POSTIVE LID_WITHIN TESTLET - PISA 2000 Mathematics) presents the released items along with citations to the items' sources and details of the item characteristics. Arguably, to answer all five pairs of items students needed to, in various degrees, refer to information in the testlets' introductions. However, there also appears to be a consistency in so much as the LID flagged pairs of items test the same mathematical skill. For example, for a pair of items from testlet M159 "Racing Car", the ability to read graphs is essential for both items. Similarly for both items from testlet M124 "Walking", skill in solving linear equations could also induce dependency. These common characteristics are in addition to both items being based on a common stimulus. For the pair of items in testlet M136 "Apples", ability in working with quadratic equations would be crucial in correctly responding to both items. However, for this specific pair, it could also be argued that item M136Q02 is serving as a crucial stimulus for item M136Q03 by revealing the mathematical expressions pointing to a quadratic equation problem.

Positive LID between pairs of items from different testlets

In PISA 2000 two pairs of items (M159Q05 with M192Q01T and M159Q05 with M266Q01T) indicated positive LID between mathematical questions from various testlets. For convenience, items from the second pair are available as reported in the electronic appendix (Electronic Appendix for Figure 5.4.2_POSTIVE LID_BETWEEN TESTLETS - PISA 2000 Mathematics). Both items require the skill of estimating the perimeter of an irregular geometric shape. This specific skill is likely to underpin the positive LID as other

item dimensions such as question format, concept or context do not match. However, both items aim to evaluate "Connections and integration for problem-solving" competency class as listed in PISA's assessment framework (OECD, 1999).

Negative LID between pairs of items from different testlets

Here, the characteristics of pairs of items located in different testlets showing negative dual-index LID (i.e. RCs ≤ -0.1 and MIs ≥ 10) are examined qualitatively. Figure 5.4.2 representing PISA LID dependency is somewhat atypical when compared to the corresponding 14 figures for other PISA waves and cognitive domains, which are progressively introduced in sections below. Many pairs of items indicate the dual-index negative LID (Electronic Appendix for Figure 5.4.2_NEGATIVE LID_BETWEEN TESTETS - PISA 2000 Mathematics). The majority of item pairs showing negative LID are items from testlets M124 and M136, paired with items from testlets M150, M155, and M159. The complete database with background information for the mathematical items allows speculation of the possible reasons for this. All items from M124 and M136 are of "Open Constructed Response" type while items from testlets M150, M155, and M159 are either of "Closed Constructed Response" or "Multiple choice" format. Questions from M124 and M136 also appear to be more difficult and frequently matched with easier questions. Perhaps students decided to be selective in their time and effort allocation which, as suggested by Yen (1993), can result in negative dependency. These and other characteristics that appear to be related to negative LID are considered below for other waves.

PISA 2003

In 2003 mathematics was a targeted cognitive domain with approximately three times (84 items) as many mathematical questions used when compared to the PISA 2000 assessment (31 items). Also, the distribution of types of testlets in regard to their size changed. As can be seen in Figure 4.3.1, PISA 2003 used 35 single item testlets compared to only eight employed in PISA 2000. At the same time, the number of four-item testlets was reduced from 3 to 2 between these two PISA waves. Figure 5.4.3 (and its electronic equivalent - Electronic Figure 5.4.3 - Mathematics PISA 2003) indicates that, as in PISA 2000, pairs of items flagging positive LID within and between testlets were found, with only a few item pairs showing negative LID.

Green lines – positive residual correlations
Red lines –negative residual correlations

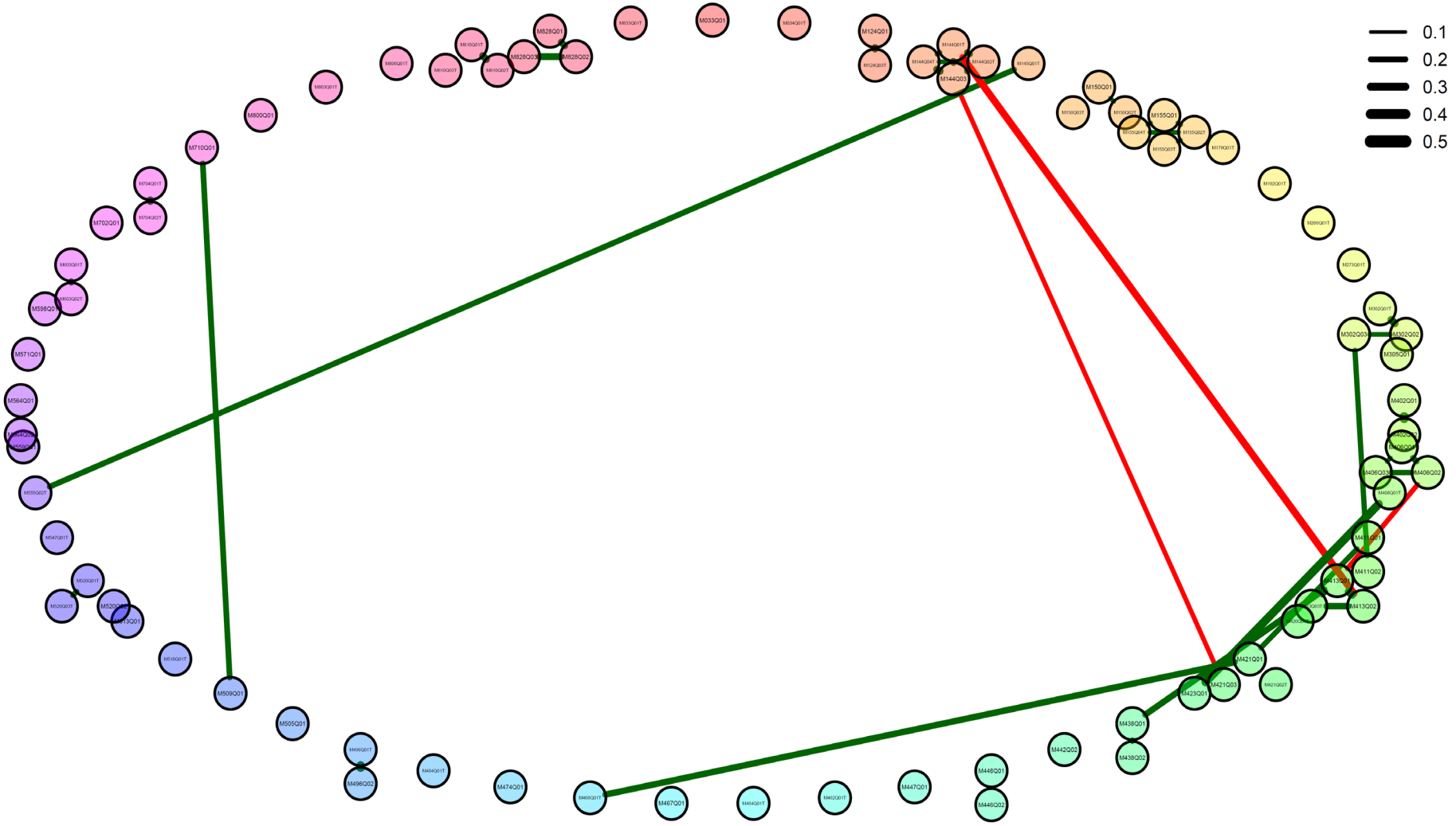


Figure 5.4.3 Visualisation of residual correlations data as a network - Mathematics PISA 2003[24]

[24] This figure has high resolution and can be zoomed in or alternatively viewed in a corresponding electronic appendix.

Positive LID between pairs of items within the same testlets

Out of all nineteen non-singular testlets, fifteen had at least one pair of questions with dual-index LID located within the testlets. In total, there were 29 pairs of such items. Electronic Appendix for Figure 5.4.3_POSTIVE LID_WITHIN TESTLET - PISA 2003 Mathematic lists all 29 pairs along with reproduced item contents for item pairs from the released testlets.

The undisclosed four-item testlet M144 "Cube Painting", which was re-used from the previous PISA 2000 wave for the purpose of linking, once again produced six LID pairs among its items. Another non-released linking testlet, M155 "Population Pyramids", generated four LID pairs. Partial information about this testlet is available in OECD (1999) which suggests that the common stimulus for its items is likely to be a collection of four graphs showing observed and predicted age distributions of the Netherlands population. Interestingly, only one pair of items from this testlet showed positive LID in PISA 2000, but this testlet's items were involved in negative LID in this first wave of PISA study. The presence of positive within-testlet dual-index LID was also duplicated for another two pairs of testlets used in the previous wave; M124 "Walking" and M150 "Growing Up". These two pairs involve released items, and the electronic appendix (Electronic Appendix for Figure 5.4.3_POSTIVE LID_WITHIN TESTLET - PISA 2003 Mathematics) shows the item descriptions. The positive LID between two items from M124 "Walking" does require respondents to consult the testlet introduction, but both items also require similar mathematical skills to solve them. Furthermore, both items use an open constructed response format requiring students to show their work. Thus, the LID observed for this testlet cannot be attributed only to the use of a common stimulus.

Testlets M302 "Car Drive", M402 "Internet Relay Chat", M413 "Exchange Rate", M438 "Exports", M520 "Skateboard", M704 "The Best Car" and M810 "Bicycles" were new to PISA 2003 but have been released. These testlets generated an additional ten, within-testlet positive LID item pairs that are reproduced in the electronic appendix mentioned above. The LID observed among the M302 "Car Drive" items appears to be driven in part by the graph in the testlet introduction, but LID could also be dictated by the common mathematical skill of reading graphs needed for all three M302 items. Similarly, graph interpretation is crucial for both items in M438. LID in a pair of items from M402 is most likely to be caused by a common prompt as both questions refer to and require referencing time delay clocks from the

testlet introduction. Common stimuli could be argued to be the predominant drivers of LID for item pairs from M810, M704 and M520. On the other hand, positive LID among three pairs of items from M413 "Exchange rate" cannot be caused by a shared prompt as an introduction in this testlet is very short and non-consequential for subsequent questions. Item-chaining could be excluded as a possible cause because each item contains all the information required to answer it correctly. It can be argued that having skills in calculating currency exchange could produce positive LID for the item pairs of this testlet. This testlet is of particular interest in the section of this research dealing with LID at the national level (see Section 6.3.1) The majority of Asian or Eastern European countries in 2003 had their own currencies and therefore currency conversion would have been common, while most Western European countries had adopted the Euro and PISA participants from those countries would not have been exposed to currency exchange to the same extent.

LID is observed between items from four other testlets new to PISA in 2003, but these have not been released. Testlet M603 "Number Check" has not been made public, a reference by Ruddock, Clausen-May, Purple, and Ager (2006) provides a plausible explanation for the observed positive LID. The authors elaborate that the M603 items, on one hand, require simple mathematical skills, but on the other hand, demand considerable language comprehension skills. Composed of three items, testlet M406 "Running tracks" reveals LID between all combinations of its items. While the unknown prompt may be causing the LID, interestingly all three questions in this testlet were of an "Open Constructed Response" item format. A pair of items from M496 "Cash Withdrawal" showed dual-index LID with a strong RC value of 0.3. Similarly, two pairs of items from M828 "Carbon Dioxide" reported RCs exceeding 0.2.

Positive LID between pairs of items from different testlets

Seven pairs of between-testlet items showed positive LID, and detailed information regarding them can be found in the electronic appendix (Electronic Appendix for Figure 5.4.3_POSTIVE LID_BETWEEN TESTLETS - PISA 2003 Mathematics). The highest RC value (.31) of all seven item pairs was for a pair M408Q01T "Lotteries" and M423Q01 "Tossing Coins". While these items are not part of the released collection, the common underlying mathematical concept of "Uncertainly and data" suggests that related mathematical knowledge and skill may be a likely reason for the positive LID.

The second, third, fourth and fifth highest LID pairs of between-testlet items involve items that were released and are subsequently reproduced in the corresponding electronic appendix (Electronic Appendix for Figure 5.4.3_POSTIVE LID_BETWEEN TESTLETS - PISA 2003 Mathematics). Positive LID for the pair M421Q01 from "Height" testlet and M468Q01T from "Science Tests" testlet is likely to arise from the specific knowledge of how to calculate an average. A similar explanation is likely for M509Q01 "Earthquake", and M710Q01 "Forecast of Rain" item pair as both questions require unique knowledge of frequentist probability. Another pair of items showing LID (M145Q01T "Cubes" and M555Q02T "Number cubes") requires the same geometrical ability to project a 3D object onto 2D explicitly referring to the same geometrical object, namely dice. A plausible explanation for LID in item pair M413Q01 and M438Q01 is less apparent. According to the PISA 2003 Technical Manual (OECD, 2005b, p. 17) both items were present in Booklet number 2 and located in clusters in the middle of testing time. Yen (1993) suggested fatigue or speediness as a possible cause of LID, but this seems unlikely given their locations in the booklet. Both items were very easy to answer as 80%, and 79% of participants from PISA 2003 international sample answered these items correctly, and both items were grouped into the "Reproduction" aspect of item competency characteristics. It is possible that the common content (Reproduction) led to the positive LID.

The final two pairs (M302Q03/M411Q02 and M411Q01/M421Q01) of between-testlet items showing positive LID both include unreleased items from testlet M411 "Driving". Both matching and released items M302Q03 "Car drive" and M421Q01 "Height" involve the concept of averages. Potentially items from M411 may also require knowledge of mean calculation. Also, the first pair of items come from testlets involving driving as suggested by their titles which may indicate some conceptual commonality of the items.

Negative LID between pairs of items from different testlets

Three pairs of items are reported with considerable negative LID. The electronic appendix (Electronic Appendix for Figure 5.4.3_NEGATIVE LID_BETWEEN TESTLETS - PISA 2003 Mathematics) shows all available item level information. Given that only one item of each pair was released no practical explanation for the LID is possible.

PISA 2006

Mathematics in PISA 2006 was not the targeted domain in PISA 2006, so no new items

were included in this wave, and all testlets were re-used from the two preceding waves of the study. Figure 5.4.4 (and Electronic Figure 5.4.4 - Mathematics PISA 2006) graphically depict pairs of items with dual-index LID.

Green lines – positive residual correlations
Red lines –negative residual correlations

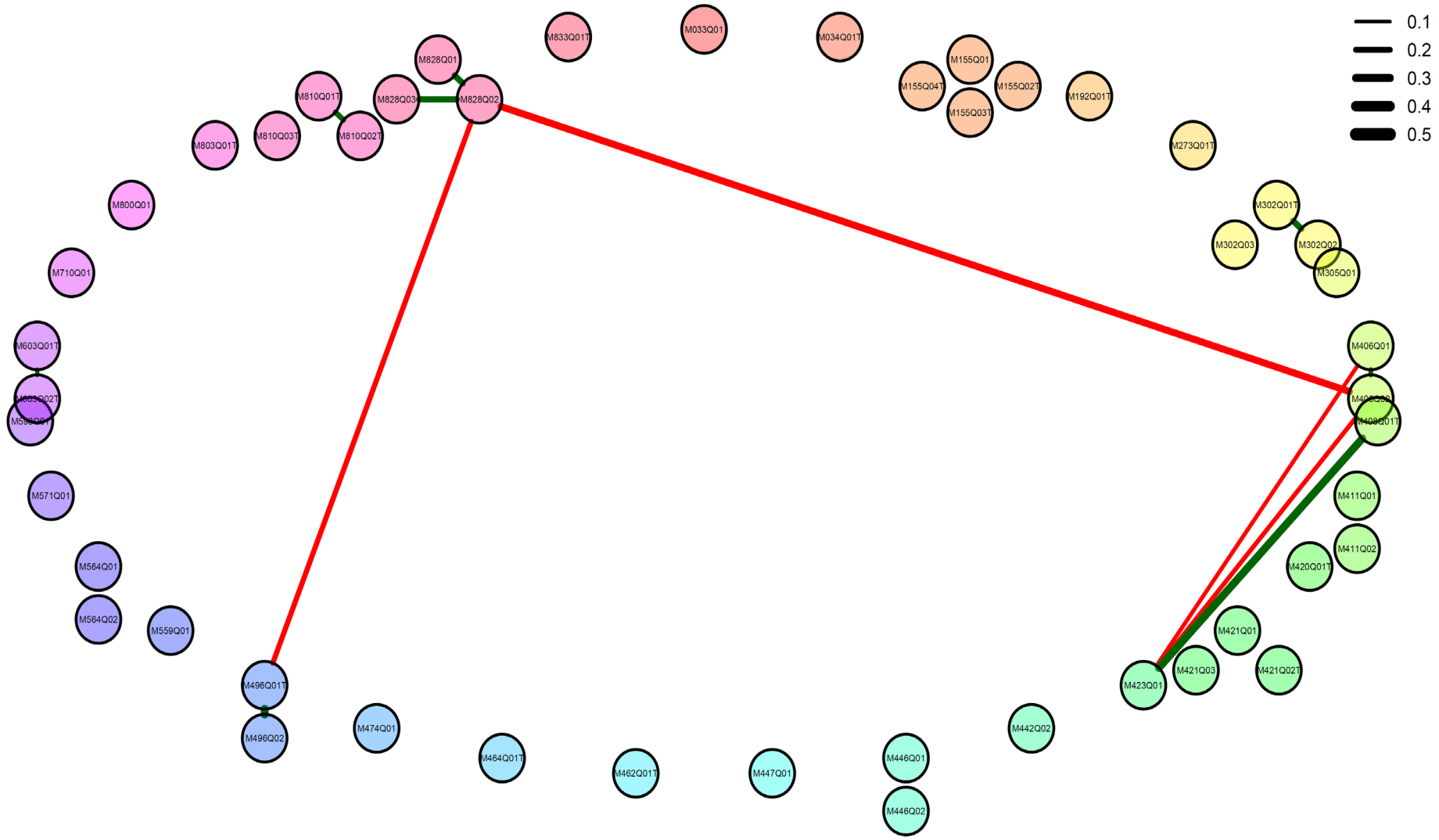


Figure 5.4.4 Visualisation of residual correlations data as a network - Mathematics PISA 2006[25]

[25] This figure has high resolution and can be zoomed in or alternatively viewed in a corresponding electronic appendix.

Positive LID between pairs of items within the same testlets

Out of eleven non-singular testlets, six had at least one pair of items showing positive LID between items within the same testlets. Electronic appendix (Electronic Appendix for Figure 5.4.4_POSTIVE LID_WITHIN TESTLET - PISA 2006 Mathematics) corresponds to the results shown in the figure above. All seven combinations of within-testlet items that showed positive LID in PISA 2006 did so also in the previous wave. Items M810Q01T and M810Q02T from the "Bicycles" testlet require consulting a table that indicates the relationship between distance travelled and a number of wheel rotations. Similarly, items M302Q01T and M302Q02 involve referring to a graph, but both require the mathematical ability to interpret graphs.

Positive LID between pairs of items from different testlets

Only one pair of items (see M408Q01T M423Q01 in Electronic Appendix for Figure 5.4.4_POSTIVE LID_BETWEEN TESTLETS - PISA 2006 Mathematics) revealed considerable positive LID (MI=24 and RC=0.29). This same pair of questions from "Lotteries" and "Tossing Coins" was also prominent in the previously reported PISA 2003, with dependency likely to be related to the common mathematical concept of "Uncertainty and data".

Negative LID between pairs of items from different testlets

Four pairs of items flagged considerable negative LID as shown in Electronic Appendix for Figure 5.4.4_NEGATIVE LID_BETWEEN TESTLETS - PISA 2006 Mathematics. Because none of the eight items involved were released, speculation about reasons for this negative LID is limited. However, two pairs (M406Q02/M423Q01 and M406Q01/M423Q01) involve difficult and "Open constructed response" type items from M406 "Running Tracks" matched with the very easy M423Q01. Furthermore, both testlets M406 and M423 were located in the same cluster (M4) and therefore were adjacent to each other. Selective time and effort allocation could be at play and according to Yen (1993) would result in negative LID.

PISA 2009

Once again in PISA 2009 mathematics was not the main tested cognitive domain. Furthermore, the number of mathematical items used was reduced by thirteen compared to PISA 2006. The reduction comprised three non-singular testlets which can be observed in

Figure 4.3.1. Figure 5.4.5 (and Electronic Figure 5.4.5 - Mathematics PISA 2009) shows positive and negative LID between pairs of items, within and between testlets.

Green lines – positive residual correlations
Red lines –negative residual correlations

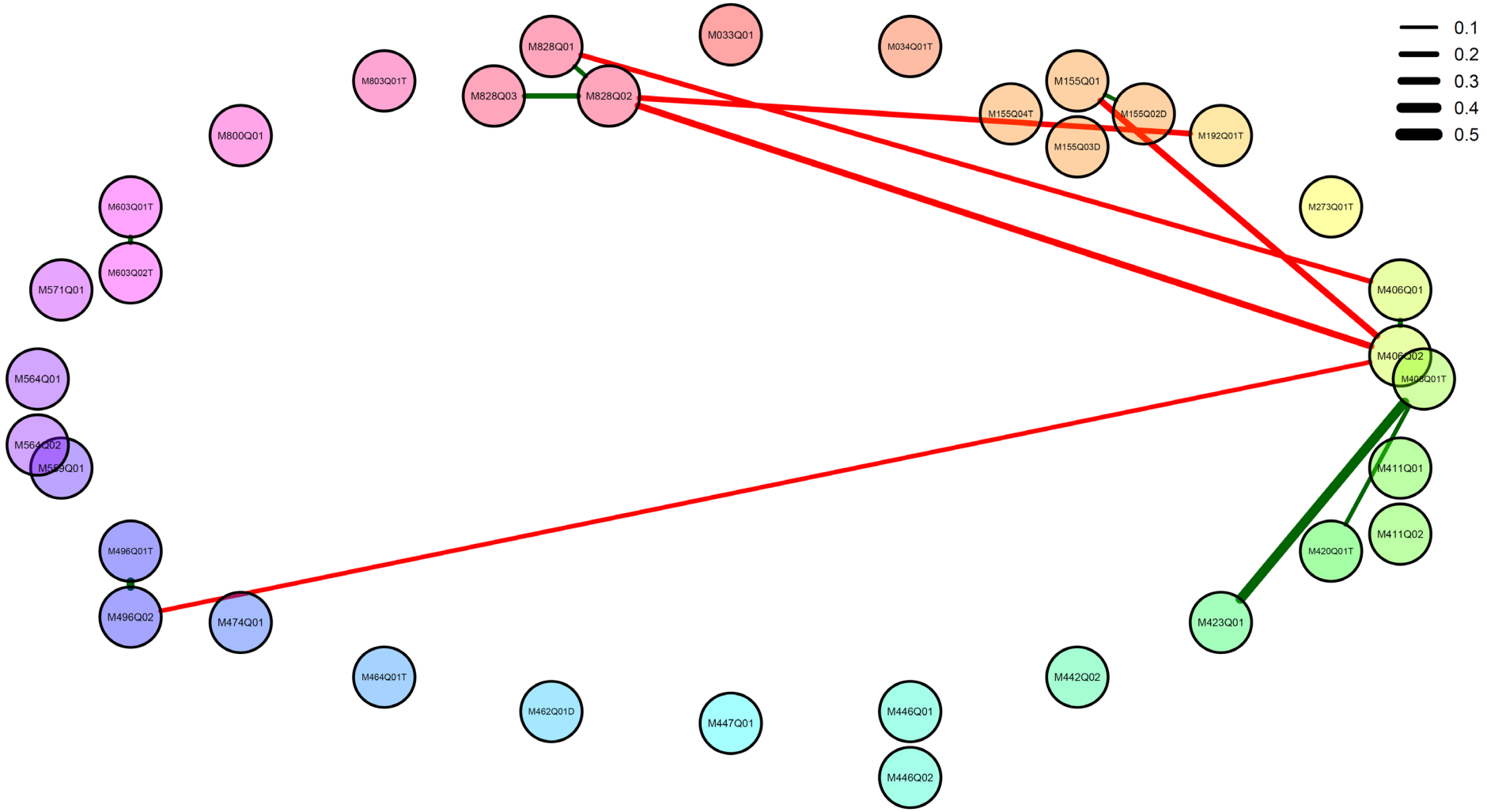


Figure 5.4.5 Visualisation of residual correlations data as a network - Mathematics PISA 2009[26]

[26] This figure has high resolution and can be zoomed in or alternatively viewed in a corresponding electronic appendix.

Positive LID between pairs of items within the same testlets

Six pairs of within-testlet questions revealed considerable positive LID as reported in the electronic appendix (Electronic Appendix for Figure 5.4.5_POSTIVE LID_WITHIN TESTLET - PISA 2009 Mathematics). Unfortunately, none of the pairs involved released items. However, all six pairs were also reported in PISA 2006 and PISA 2003 as having positive LID. This cross-wave consistency is worth highlighting given that different cohorts of 15 years old students are involved in each PISA data collection and in most countries different schools are sampled from wave to wave.

Positive LID between pairs of items from different testlets

Electronic appendix (Electronic Appendix for Figure 5.4.5_POSTIVE LID_BETWEEN TESTLETS - PISA 2009 Mathematics) reports two pairs of items from different testlets with positive LID. Since their introduction in 2003, the pair M408Q01T "Lotteries" and M423Q01 "Tossing Coins" consistently produced high LID with the RC value on this occasion being 0.39. Item M408Q01T also show signs of LID with M420Q01T "Transport" which, while not released, also tests the "Uncertainty and data" mathematical concept. In the previous two waves, this pair of items showed no signs of LID. There were also four pairs of items with MIs exceeding ten but RCs above 0.9, only slightly short of being labelled dual-index LID. With none of eight items being released, limited interpretation can be offered. However, descriptive exploration of various items characteristics highlighted that all pairs of items belonged to the same clusters representing 30 mins of students' testing.

Negative LID between pairs of items from different testlets

Five pairs of items showed considerable negative LID (Electronic Appendix for Figure 5.4.5_NEGATIVE LID_BETWEEN TESTLETS - PISA 2009 Mathematics). The interpretation is not possible as none of the items were released. However, four of the pairs involve items from testlet M406 "Running Tracks". Furthermore, three other negative LID pairs with MI exceeding 10 and only marginally below the dual-index LID cut point, with RCs around -0.9, involve items from M406.

Items from this testlet (M406) contributed to negative LID in two out of three pairs in PISA 2006. From the information that is available about these items, it can be seen that both were quite challenging and in PISA 2009, only about 17% and 27% of students responded to

them correctly. Both are of the "Open Constructed Response" item type. Thus, it is possible that selective allocation of effort, as suggested by Yen (1993) may be a factor.

PISA 2012

Mathematics was the major domain assessed in PISA 2012. The number of items increased to 84[27], almost tripling the number from the previous PISA 2009 wave. The majority of new items for PISA 2012 evaluated in this chapter came in the form of two-item testlets (7 new testlets) and three-item testlets (9 new testlets). Figure 5.4.6 (and Electronic Figure 5.4.6 - Mathematics PISA 2012) offers a visual overview of LID between pairs of items.

[27] In PISA 2012 109 mathematical items were used. As this chapter only looks at OECD countries, items that were used in the 2012 in the 'easy booklets' calibration are not included.

Green lines – positive residual correlations
Red lines –negative residual correlations

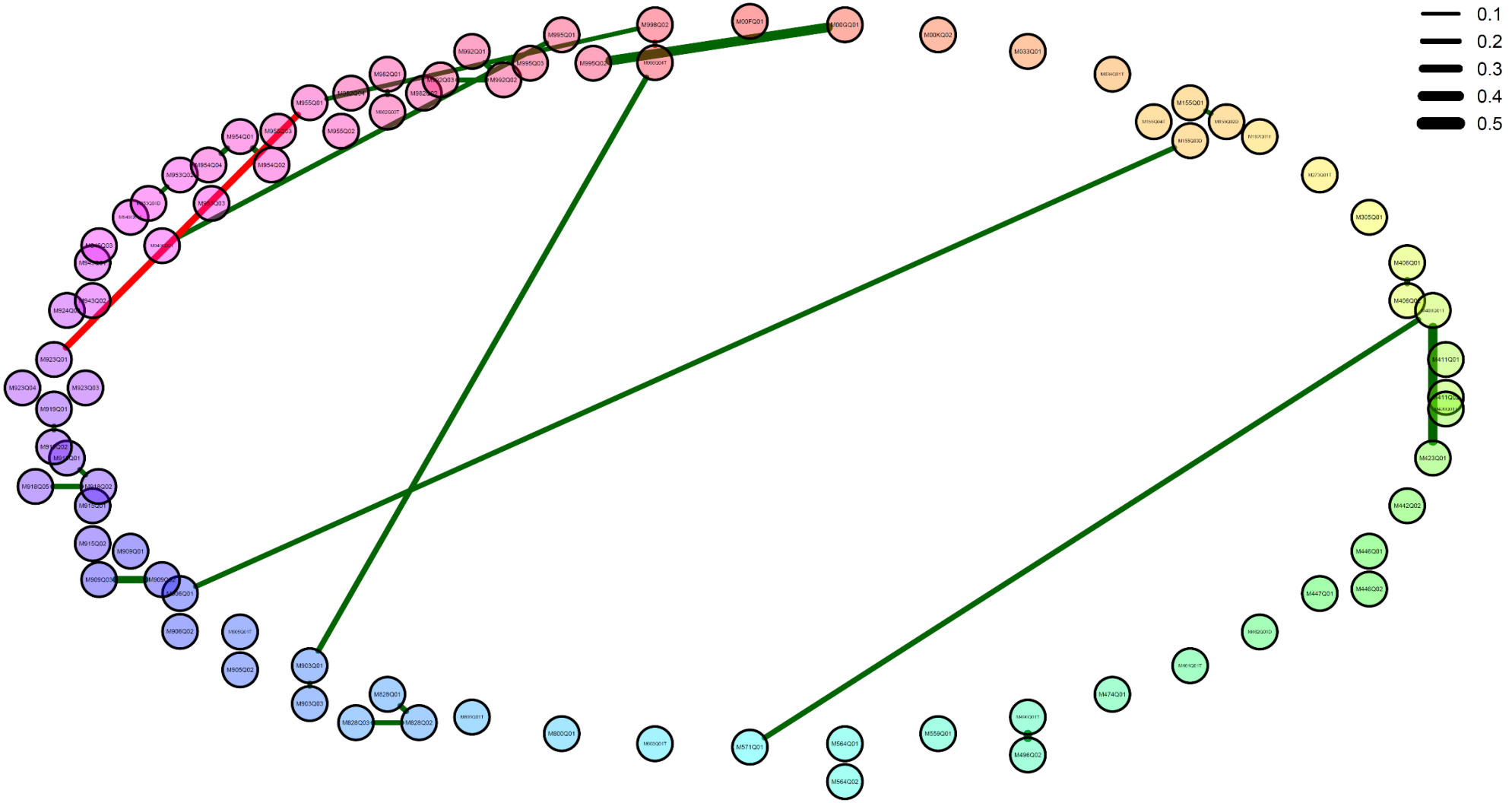


Figure 5.4.6 Visualisation of residual correlations data as a network - Mathematics PISA 2012[28]

[28] This figure has high resolution and can be zoomed in or alternatively viewed in a corresponding electronic appendix.

Positive LID between pairs of items within the same testlets

Sixteen pairs of mathematical items (Electronic Appendix for Figure 5.4.6_POSTIVE LID_WITHIN TESTLET - PISA 2012 Mathematics) indicated positive LID among items belonging to the same testlets. Half of the non-singular testlets had at least one pair of items with dual-index LID. With a large number of new items in PISA 2012 not having been released, only three pairs can be viewed and evaluated for their possible causes of LID.

Two released items from M903 "Drip rate" clearly required a common reference to the opening passage for the drip rate formula. However, common to both items is the ability to manipulate equations, and this seems likely also to contribute to LID. Similarly, three items from M918 "Charts" contribute to two LID pairs and have dual LID origin with both common stimuli for both items and a common skill of graph reading also required for both. Five out of six PISA 2009 within-testlet pairs which showed positive LID also are prominent for PISA 2012. The remaining eight pairs with non-released items involve two items for testlets: M909 "Speeding fines", M919 "Zs Fan Merchandise", M953 "Flu Test" and M982 "Employment Data" and three items for M954 "Medicine Doses" and M992 "Spacers". Once again four pairs of items (M155Q01 and M155Q04T, M923Q01 and M923Q04, M953Q03 and M953Q04D, M949Q01T and M949Q02T) have RCs and MIs very close to the dual-index LID cut point. The pair from M155 shows a similar tendency as in previous five waves of PISA.

A pair of items from M923 "Sailing Ships" were only marginally short of making a dual-index LID rule. They were released and are worth mentioning as a common introduction for this pair is of no consequence, but the underlying skill of percentage usage is common to both items and is a likely cause of positive LID.

Positive LID between pairs of items from different testlets

Once again pair M408Q01T M423Q01 produced the second largest RC (Electronic Appendix for Figure 5.4.6_POSTIVE LID_BETWEEN TESTLETS - PISA 2012 Mathematics) between items from different testlets. This result is consistent with corresponding sections for PISA 2003, 2006 and 2009. For the 2012 assessment, the largest RC of 0.38 occurred between two new items M00GQ01 from M00G "An Advertising

Column" and M995Q02 from "Revolving Door". Unfortunately, only one of the items has been released. However, it is known that both items were very difficult with only 9% and 3.5%, respectively of international sample students providing correct responses. Both items also cover the "Space and shape" item concept and were both allocated to the "Formulate" item process category. Another item, M995Q01 from the "Revolving door" testlet showed positive LID with the undisclosed item M949Q02T "Roof Truss Design". Both items were from "Space and shape", and the title of the M949 testlet offers a possibility that maybe both items require knowledge of angles. The LID for the remaining four pairs of items M903Q01 and M998Q04T, M155Q03D and M906Q01, M955Q01 and M998Q02, and M408Q01T and M571Q01 cannot be explained as they involve non-released items.

Negative LID between pairs of items within the same testlets

In this whole LID investigation, there was only one pair of items within the same testlet that produced considerable negative LID (Electronic Appendix for Figure 5.4.6_NEGATIVE LID_WITHIN TESTLET - PISA 2012 Mathematics**)**. The items involved were introduced in PISA 2012 and are part of testlet M998 called "Bike rental". This testlet has not been released. However, this is the only one testlet for which Israel was identified as the source, so perhaps some translational or specific cultural or curriculum feature contributed to the presence of negative LID.

Negative LID between pairs of items from different testlets

Only one pair of items (M923Q01 and M955Q01) showed in PISA 2012 dual-index LID (Electronic Appendix for Figure 5.4.6_NEGATIVE LID_BETWEEN TESTLETS - PISA 2012 Mathematics). As testlet M955 is not released, no practical explanation can be offered.

5.4.1.2 Summary and cross wave consistency of LID in the mathematics domain

Positive LID between pairs of items within the same testlets

The overall conclusion about the factors contributing to positive dependency among items from the same testlets is that it is unlikely to be only due to common stimuli expressed by text or graph. A review of released items suggests that matching mathematical skills are likely to be partially responsible for LID with some testlets such as M413 "Exchange Rate" having a negligible stimulus. Item chaining was also found to be a plausible positive dependency

driver for some items, for example pair M136Q02 and M136Q03. As the majority of released items come from the initial two waves of PISA, explanations for dependency are limited for PISA 2009 and 2012. However, the systematic approach to reporting along with comprehensive results being available in the electronic appendices, should facilitate further in-depth investigations for readers with full access to all mathematical items. Additional insides into presence of positive LID are presented in Section 5.4.2.1 reporting quantitative results.

Mathematics dual-index dependent results, for within-testlet positive LID, produced cross wave consistency. To give an additional point of view and to facilitate a cross-wave overview, Electronic Appendix for Figures 5.4.2-6 offers a listing of all dual-index LID producing mathematical items reported in Section 5.4.1 cross tabulated against PISA wave. The non-released four-item testlet M144 "Cube Painting" showed positive LID for all six pairs of its items on both occasions when it featured in PISA 2000 and 2003. The publication by Monseur et al. (2011) examining mathematical dependency in PISA 2003 featured the M144 testlet as showing pairwise LID but being marginally short of qualifying as 'global context dependence'. This term used by the authors aims to represent a 'whole testlet LID'. The results of this research would put M144 testlet as presenting a global context dependence. Monseur et al. (2011) showed that the two-item M124 "Walking", the two-item M496 "Cash Withdrawal" and the three-item M828 "Carbon Dioxide" testlets produced dependency. These results were confirmed in the sections above, but also added conclusions that positive LID for these testlets was consistently present in all PISA studies when M124, M496 and M828 were used. Testlet M124 was utilised twice in PISA 2000 and PISA 2003 while M496 and M828 were used four times in 2003, 2006, 2009 and 2012. Results for testlets M402 "Internet Relay Chat" and M413 "Exchange Rate", which were used only in PISA 2003, revealed positive LID, as was shown by Monseur et al. (2011). Similarly, results for M406 "Running Tracks" and a single pair from M810 "Bicycles" agreed between both publications in regard to PISA 2003 with a dependency between two of its M406 items (M406Q01 and M406Q02) extending to the later three implementations of the PISA. Furthermore, pair M810Q01T and M810Q02T presented cross wave consistency in dependence also for PISA 2006. Interestingly two-item testlet M603 "Number Check" which in this study showed positive LID in three PISA waves was not identified by Monseur et al. (2011) using the PISA 2003 data. While this testlet has not been released, Ruddock et al. (2006), suggest that two items from M603 require in-depth language comprehension, but that

is not helpful in explaining the lack of cross-research consistency. The results of within-testlet dependency among mathematical items in PISA 2000 are also partially confirmed by Cai (Cai, 2010; Cai et al., 2011) who report better fitting models when testlets are controlled for. Both of Cai's publications used subsets of PISA 2000, and all testlets employed by him feature in this study, with at least one pair of within-testlet items showing positive LID. It is worth highlighting that some of the testlets introduced in PISA 2012 also show positive LID among all of their items, e.g. M992 "Spacers", M954 "Medicine Doses" and M918 "Charts". Publication by Koğar and Kelecioğlu (2017) used PISA 2012 mathematical datasets and confirmed that statistical models factoring testlets performed better. Authors mentioned explicitly in their discussion about few pairs of items reporting high levels of item dependency, but they did not list them limiting the possibility to cross-validate with this study results.

In conclusion, despite the limited number of publications investigating within-testlet dependency of PISA's mathematical items, the results of this study largely concur with Monseur et al. (2011), Cai (2010) and Cai et al. (2011). However, as the analyses were undertaken in five PISA waves, it is evident that some of the testlets used for the purpose of cross-wave linking showed consistency in flagging dependency in multiple PISA studies in which they were utilised.

Positive LID between pairs of items from different testlets

A literature search failed to locate a publication that investigates the dependency among PISA items from different testlets. The results presented in Section 5.4.1 show that this type of dependency is present and due to the analyses of released items, logical explanations for its existence can be provided. Items' pair featuring M408Q01T from M408 "Lotteries" and M423Q01 from M423 "Tossing Coins" reported residual correlations of about 0.3 in all four PISA assessments in which the items were used. Although, neither of them are released to the public the testlets' titles and common content strand of "Uncertainty and data" suggest a plausible reason for dependency. At the same time, there were item pairs featuring in more than one study that showed positive LID on only one occasion (see also Electronic Appendix for Figures 5.4.2-6). The strongest example of this came from M408Q01 and M420Q01 revealing LID only in PISA 2009. The majority of the between-testlet positively dependent item pairs were used only in one PISA study, and therefore did not permit an evaluation of cross-wave consistency of LID. As reported above, it was found that specific skills of

estimating the perimeter of an irregular geometric shape, ability to calculate an average or having the skills to project a 3D object onto 2D were likely causes for between-testlets LID (based on the use of two LID indicators, cut point of RCs≥0.1 and MIs≥10). The usage of dual-index LID, while driven by pragmatic reasons to make qualitative in-depth investigations manageable, produced a limitation which needs to be acknowledged. As reported in Figure 5.4.1 there were pairs of between testlet items with high positive residual correlations which failed to fulfil the requirement of high modification indices. The qualitative investigation does not look at these item pairs. However, a quantitative component of this chapter includes them as LID indicative for the purpose of logistic regressions.

Negative LID

In this research only the mathematics domain revealed within-testlet items with dual-index negative dependency, and only one such pair was identified. The pair (M998Q02 and M998Q04) came from mathematical testlet M998 "Bike Rental" introduced in 2012. As this testlet is not available for the public to view, it was not possible to pursue plausible reasons for its LID. No apparent dual-index negative LID cross-wave patterns were present as can be seen in the electronic appendix (Electronic Appendix for Figures 5.4.2-6) summarising all the mathematical qualitative results. However, for one pair (M406Q02 and M828Q02) negative dependency was featured twice. This was not entirely cross-wave consistent as on two other occasions when both items were used, the negative LID was not found. Speculative reasons for negative LID were proposed pointing to Yen's (1993) suggestion that selective time and effort allocation can produce it. However negative item dependency may also be an artefact of the existence of positive LID as implied by Habing and Roussos (2003). As mentioned in a previous paragraph, the dual-index LID rule of thumb also proved to be too restrictive for some item pairs with residual correlations less than -0.1 but not fulfilling the modification indices cut point. This limitation is acknowledged in the corresponding section 7.2.1.

5.4.1.3 Qualitative investigation of reasons for LID in the reading domain

As was the case with the mathematics domain, the qualitative investigation of dual-index LID will follow the same pattern of subsections, discussing each PISA implementation with separate paragraphs to elaborate on positive and negative LID as well as its within- or between-testlet allocation. Once again, network qgraph plots are featured for all five PISA waves along with electronic appendices reproducing all the items which were released to the public.

PISA 2000

Reading was a primary targeted domain in the inaugural PISA study. Thirty-seven testlets were used in total with only two of them involving single questions. Out of 35 non-singular testlets, 69% had at least one pair of within-testlet items with dual-index positive LID. Eight pairs of items from non-matching testlets indicated positive dual-index LID and ten reported considerable negative LID. Figure 5.4.7, and its pdf equivalent provided for in-depth viewing (Electronic Figure 5.4.7 - Reading PISA 2000) gives a graphical overview of item pairs showing dual-index LID.

Green lines – positive residual correlations
Red lines –negative residual correlations

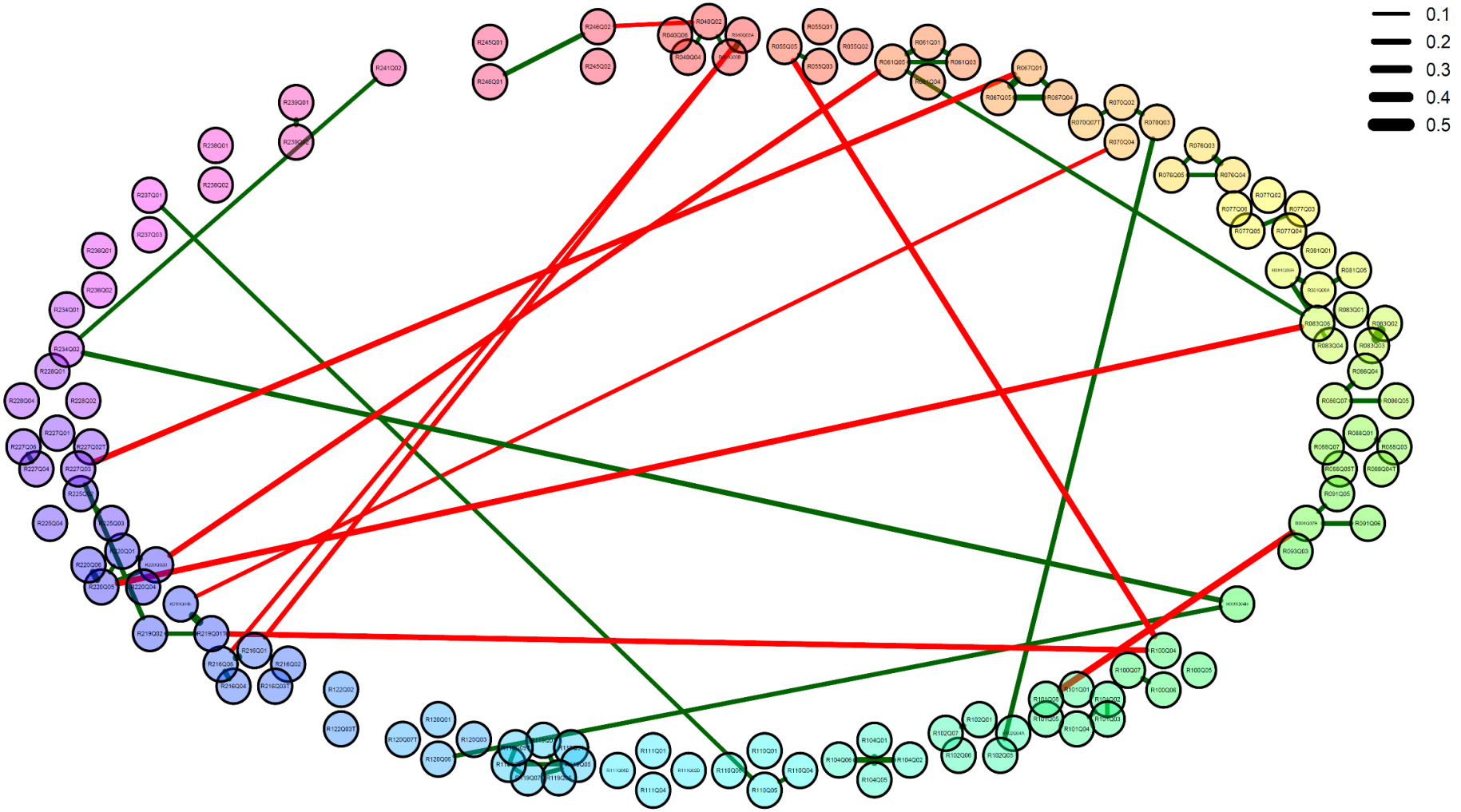


Figure 5.4.7 Visualisation of residual correlations data as a network - Reading PISA 2000[29]

[29] This figure has high resolution and can be zoomed in or alternatively viewed in a corresponding electronic appendix.

Positive LID between pairs of items within the same testlets

In total 49 pairs of items produced positive dual-index LID within the testlets. Six testlets (R040 – "Lake Chad", R061 – "Macondo", R067 – "Aesop", R076 – "Iran Air", R119 – "Gift" and R220 – "South Pole") had at least three of their items involved in dual-index positive LID. Out of these, R067 and R076 produced LID for all their items' combinations, and the seven-item R119 had five pairs flagging LID. Of the forty-nine pairs of items revealing LID, 21 pairs of items were released (Electronic Appendix for Figure 5.4.7_POSTIVE LID_WITHIN TESTLET - PISA 2000 Reading), which therefore facilitated a qualitative review of plausible drivers of LID.

A pair of items R040Q03A and R040Q03B from the "Lake Chad" testlet understandably produced positive LID as both items refer to the same sentence in the testlet introductory text. LID for two pairs R040Q02/R040Q04 and R040Q02/R040Q03B is likely to be caused by the need to refer to introductory figures and by the common underlying ability to read the graphical prompts. Similarly, three LID pairs of items from "Macondo" clearly refer to the second half of a piece of prose. Further, items R061Q01, R061Q03 and R061Q05 all ask about the same aspect to be inferred from the text. Interestingly, testlet R077 "Flu", that had five items in PISA 2000, produced only one dual-index LID pair of items (R077Q03 and R077Q05). LID for two pairs R040Q02/R040Q04 and R040Q02/R040Q03B is also likely to be caused by the need to refer to introductory figures and by the common underlying ability to be able to read the graphical prompts. Positive dual-index LID for items from R088 "Labour"[30] and R091 "Library" appears to be driven by graphical stimuli that are a flow chart and map for "Labour" and "Library", respectively. Another released pair showing dual-index LID (R100Q06 and R100Q07) comes from the testlet "Police", and it is likely caused by the need to refer to the testlet introductory text. The common stimulus is also likely to be responsible for positive LID between items from R110 "Runners". The most prevalent example of LID was testlet R119 "The Gift". Out of the seven items belonging to this testlet, six were part of dependency relationships, most likely induced by the lengthy introductory reading of over 1700 words. Finally, the released testlet R216 "Amanda & the Duchess" produced two pairs of items with positive LID. Of the non-released testlets, R220 "South Pole" is worth highlighting as it comprised five questions, which were involved in four pairs

[30] The source of information about R091 testlet does not explicitly use the PISA standard item labelling. It is assumed that item 7A in released item source (OECD, 2010a) is in fact Q06. This is additionally confirmed by matching simple multiple choice format.

of items with LID. While no preview of this item is available, it is known that the text type involved was a chart or map, which likely constituted a common reference on which all the items depended.

Positive LID between pairs of items from different testlets

Eight pairs of items from various testlets produced considerable positive LID as reported in the electronic appendix (Electronic Appendix for Figure 5.4.7_POSTIVE LID_BETWEEN TESTLETS - PISA 2000 Reading). Two of the pairs (R099Q04B/R234Q02 and R099Q04B/R120Q06) came from released testlets, and in both, item R099Q04B was featured. The explanation for positive LID is not immediately apparent by comparing the matching items. However, reviewing the scoring procedures for these questions provides possible explanations for dependency between the items. This item proved to be very difficult for students to answer correctly with only about 11% of participants answering correctly. After reviewing the scoring guide (National Center for Education Statistics, 2016c), the procedure for granting the fully correct score appears to be complicated, requiring a correct answer to the complimentary introductory question as well as a written response including explicit references to two facts. As one of the required facts is given in the question's statement, it is likely that many students would not repeat it in their written responses and therefore would not receive full credit. When investigating in detail matching items R234Q02 and R120Q06, the requirements regarding the scoring of these items are also somewhat confusing and require two parts to be simultaneously addressed for a complete score. To correctly answer item R234Q02 from the "Personnel" testlet, according to the scoring rubric (National Center for Education Statistics, 2016c, p. 71), the students needed to mention two ways in which the organisation described in the text was supposed to operate. However, the item requires providing two references related to a limited subgroup of people, namely those who will lose their jobs because of reorganisation. The text which is crucial to providing correct answer states that

> CIEM acts as a mediator for employees who are threatened with dismissal resulting from reorganization, and assists with finding new positions when necessary. (National Center for Education Statistics, 2016c, p. 69)

There are two ways to look at this text. Firstly, it could be treated from a perspective of two CIEM functions being separated by an intervening clause. This view is required to score this item as correct according to the scoring rubric. Secondly, the wording of this text could also be interpreted that the responsibility of finding new postitions is not explicitly limited to

a specific subgroup, i.e. employees who are threatened with dismissal. The rest of the text lists other CIEM functions suited for employees intending to voluntarily search for another job. Similarly, scoring for question R120Q06 is very elaborate (OECD, 2010a, p. 190) requiring the respondents to provide a two part answer to receive full credit. As suggested in Yen (1993, p. 190) the items that are scored in the same fashion may produce LID. This appears to be a plausible explanation of between-testlet dependency for the two examples listed above. The remaining six pairs of LID items cannot be discussed in detail because they have not been released, but four of the pairs do involve open-ended questions.

Negative LID between pairs of items from different testlets

Ten pairs of items from different testlets produced considerable negative LID (Electronic Appendix for Figure 5.4.7_NEGATIVE LID_BETWEEN TESTLETS - PISA 2000 Reading) with only two of the pairs having both items released (R040Q03A/R216Q01 and R040Q03A/R216Q06). Both pairs featured one item from the "Lake Chad" testlet and another item from "Amanda & the Duchess". The item R040Q03A was correctly answered by about 50% of the OECD sample while items from R216 testlet were easier with 74% and 67% of students answering them correctly. Both these items were preceded by the very lengthy reading of two texts totalling over 900 words. It is possible that the negative LID is driven by selective time allocation as suggested by Yen (1993) when students unexpectedly underperformed on easier items from "Amanda & the Duchess" due to the limiting time and effort needed to be invested in reading lengthy texts.

PISA 2003

Assessment of reading literacy was not a major target for PISA 2003. Consequently, only eight testlets with a total of 28 reading items were used in this wave of the study. All the items were reused from PISA 2000, although the number of items from within some of the testlets such as R102, R104, R111 and R227 were reduced. Figure 5.4.8 (Electronic Figure 5.4.8 - Reading PISA 2003) shows that there was no considerable positive LID between items from different testlets with some evidence of within-testlet positive LID and negative LID involving predominantly two testlets.

Green lines – positive residual correlations
Red lines –negative residual correlations

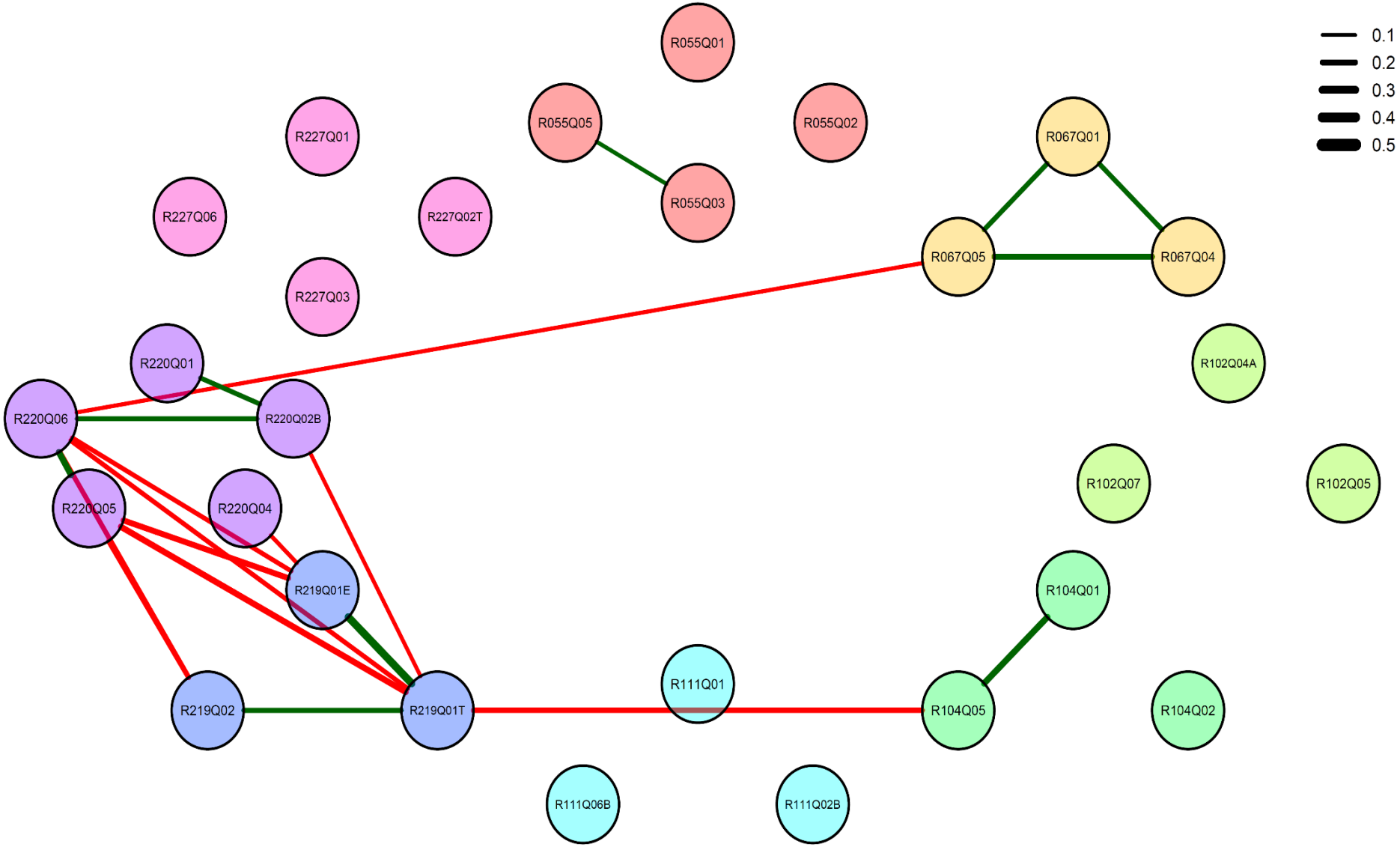


Figure 5.4.8 Visualisation of residual correlations data as a network - Reading PISA 2003

Positive LID between pairs of items within the same testlets

There were ten pairs of the linking items within the same testlets which showed considerable positive LID (Electronic Appendix for Figure 5.4.8_POSTIVE LID_WITHIN TESTLET - PISA 2003 Reading). All three items from R067 "Aesop" produced dependencies among all its items. Testlets R220 "South Pole" and R219 "Employment" also indicated positive LID among many of their items. Single pairs of LID items were featured in R055 "Drugged Spiders" and R104 "Telephone". As none of the reading items used in PISA 2003 were released, an in-depth investigation of the plausible underlying causes of LID is not possible. Dependency is likely to be related to a common introduction preceding items from the same testlets. Interestingly, nine out of ten LID pairs featured in this section about PISA 2003 flagged the same dual-index LID when investigated with PISA 2000 data.

Positive LID between pairs of items from different testlets

No positive LID was found for any between-testlet item pairs.

Negative LID between pairs of items from different testlets

Negative dual-index LIDs pairs of items were concentrated between items from two testlets R219 "Employment" and R220 "South Pole", both of which were allocated to the same cluster R1 (Electronic Appendix for Figure 5.4.8_NEGATIVE LID_BETWEEN TESTLETS - PISA 2003 Reading). All dependent items from these two testlets were of easy to moderate difficulty, with the percentage of OECD sample answering them correctly ranging from 57% to 83%. The format of the items appears to vary with all questions from R220 being of simple multiple choice format while R219 featured short response or open constructed response formats. Therefore, the negative LID may relate to the different response formats required for items in the dependent pairs.

PISA 2006

Reading was not a primarily targeted cognitive domain in PISA 2006. All the reading items used in the previous wave, 2003, were reutilised. Figure 5.4.9 (Electronic Figure 5.4.9 - Reading PISA 2006) graphically represents pairs of items with dual-index LID, and it closely mimics the dependency results from the previous PISA wave.

Green lines – positive residual correlations
Red lines –negative residual correlations

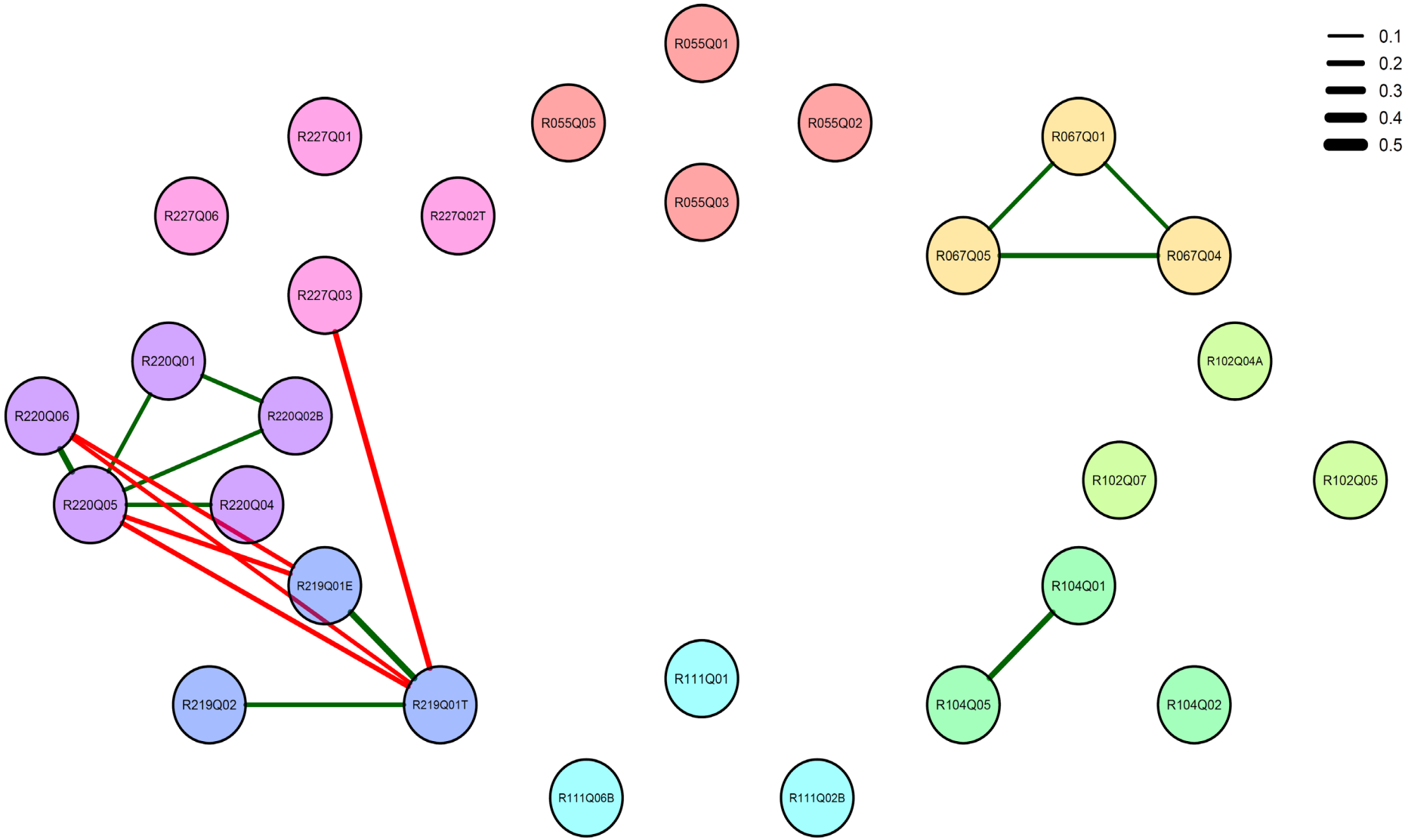


Figure 5.4.9 Visualisation of residual correlations data as a network - Reading PISA 2006

Positive LID between pairs of items within the same testlets

Eleven pairs of items showing positive LID were observed for reading in PISA 2006 (see also Electronic Appendix for Figure 5.4.9_POSTIVE LID_WITHIN TESTLET - PISA 2006 Reading). The dependency patterns for testlets R067, R104 and R219 match perfectly with the results from PISA 2003 and PISA 2000. LID in R220 also mimics the previous two PISA waves with a closer resemblance to PISA 2000. As a consequence of no PISA 2006 reading items having been released, no further discussion about the nature of the observed LID is offered.

Positive LID between pairs of items from different testlets

Similarly, in the corresponding results for PISA 2003, there were no pairs of dual-index positive LID between items from different testlets.

Negative LID between pairs of items from different testlets

Interestingly, patterns of negative LID associations point to similarities with the PISA 2003 wave, where items from testlet R219 "Employment" produced negative LID with various items from R220 "South Pole". This time five such considerable dependencies are present as reported in Electronic Appendix for Figure 5.4.9_NEGATIVE LID_BETWEEN TESTLETS - PISA 2006 Reading.

PISA 2009

Nine years after the initial PISA assessment featured reading as the main literacy of interest, reading was once again the targeted cognitive domain in 2009. Eighteen new reading testlets were featured in this assessment. For the purpose of linking across PISA waves, most reading questions contained in PISA 2006 and 2003 were reused. Some of the original PISA 2000 testlets such as R083 "Household Work", R101 "Rhinoceros", R102 "Shirts" and R245 "Movie Reviews" were also reused. Of 28 non-singular testlets, 18 had at least one pair of items showing positive within-testlet LID. Positive and negative between-testlet LID was present as shown in Figure 5.4.10 (see also Electronic Figure 5.4.10 - Reading PISA 2009).

Green lines – positive residual correlations
Red lines –negative residual correlations

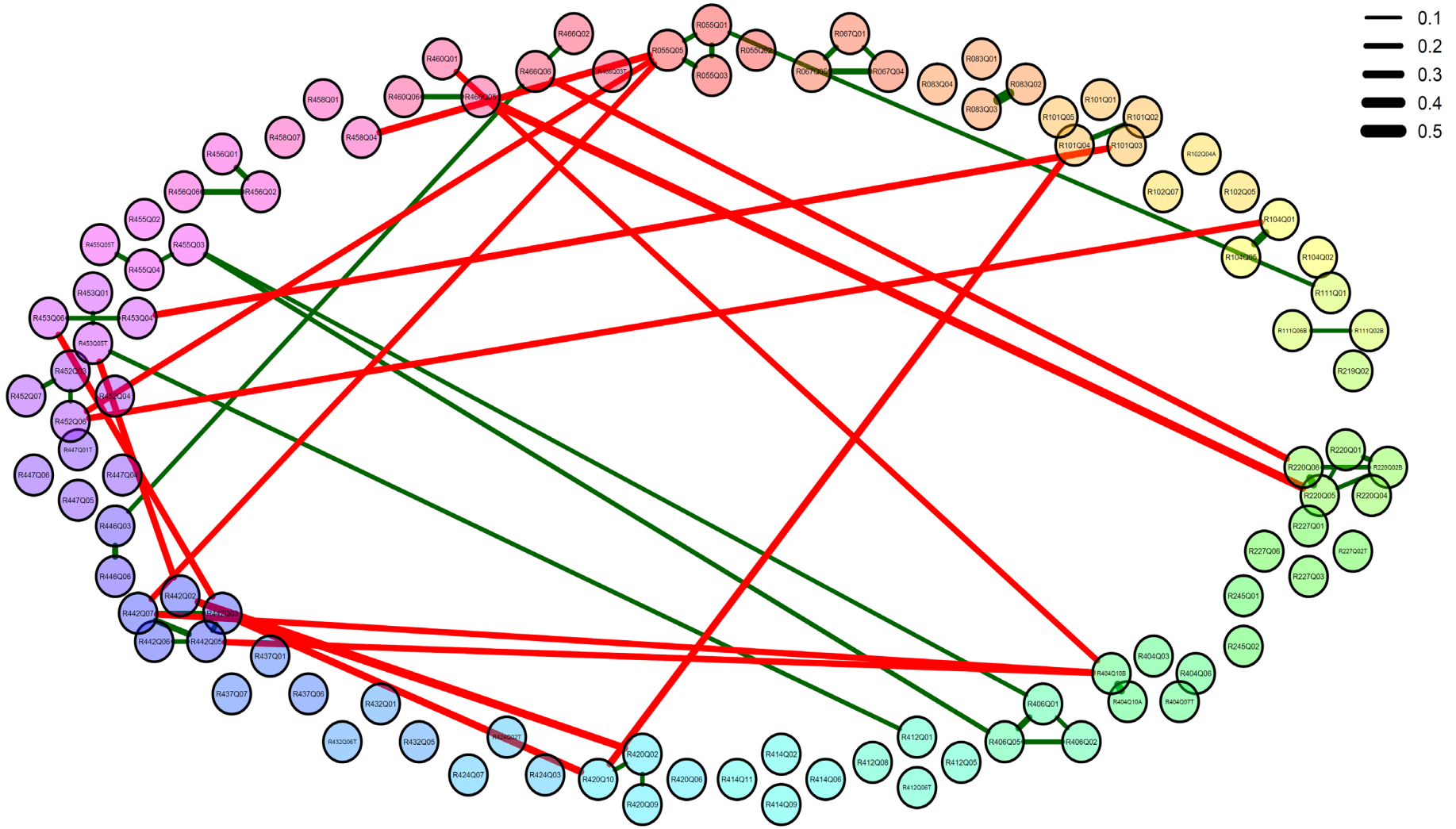


Figure 5.4.10 Visualisation of residual correlations data as a network - Reading PISA 2009[31]

[31] This figure has high resolution and can be zoomed in or alternatively viewed in a corresponding electronic appendix.

Positive LID between pairs of items within the same testlets

As mentioned in the section above, about 62% of reading testlets had at least one pair of items with considerable positive LID. In total, there were 38 pairs of items for which RC exceeded the absolute value of 0.1 and MIs were larger than 10. As the majority of the released reading items were made available soon after first the PISA assessment and could not be used again, only two pairs from R452 "The Play's the Thing" testlet are available for closer inspection (see also Electronic Appendix for Figure 5.4.10_POSTIVE LID_WITHIN TESTLET - PISA 2009 Reading). Arguably, the positive LID between items from R452 is driven by the necessity of reading and referring to the introductory text. While no additional qualitative elaboration can be produced for other testlets, it is worth highlighting a cross-wave consistency in item dependency patterns. Within-testlet positive LID between items from R067, R104 and R220 has been produced for all four implementations of PISA waves described so far. Furthermore, some of the PISA 2000 testlets that were reintroduced in PISA 2009 also present matching dependencies despite there being close to 10 years difference between cohorts for which they were used. For example, a pair of non-released items R083Q02 and R083Q03 from "Household Work", produced the highest RC of 0.4 for PISA 2009 as it did for PISA 2000. Likewise, pair R101Q02 and R101Q04 from "Rhinoceros" featured positive LID in both studies. About half the pairs of items listed in the Electronic Appendix mentioned above, come from new testlets introduced in PISA 2009. Testlet R406 "Kokeshi Dolls" yielded considerable positive LID among all three of its items. Five-item "Galileo" (R442) had five pairs of its items indicating dependency among them.

Positive LID between pairs of items from different testlets

While no between-testlet positive LID was found in the PISA 2003 and 2006 waves, five pairs of questions reveal positive between-testlet LID in the 2009 wave (see also Electronic Appendix for Figure 5.4.10_POSTIVE LID_BETWEEN TESTLETS - PISA 2009 Reading). None of the featured items are released, hindering any attempts to explain the reason behind the positive LID. Reviewing known properties of involved items such as their difficulty, item format, text format, text type, situation or aspect does not offer any striking patterns.

Negative LID between pairs of items from different testlets

Electronic Appendix for Figure 5.4.10_NEGATIVE LID_BETWEEN TESTLETS - PISA 2009 Reading lists sixteen pairs of between testlets items indicating negative LID. Given the scarcity of the released items, it is not surprising that there is not a single pair with both items that could be viewed. Section 5.4.2.3 that utilises a multilevel logistic regression gives some suggestions regarding negative dependency drivers which in the absence of released items cannot be investigated in this qualitative elaboration.

PISA 2012

In PISA 2012, reading was not the main assessed cognitive domain, and therefore the number of items and testlets used was considerably reduced, using only 44 questions in total. Only one testlet, R220, was retained from the PISA 2000 wave, with the remaining 12 testlets being introduced in PISA 2009 and reused for the purpose of linking. As in previous sections Figure 5.4.11 (see also Electronic Figure 5.4.11 - Reading PISA 2012) presents the positive LID involving items within and between testlets. Considerable negative LID is also present, mainly involving two items from testlet R404.

Green lines – positive residual correlations
Red lines –negative residual correlations

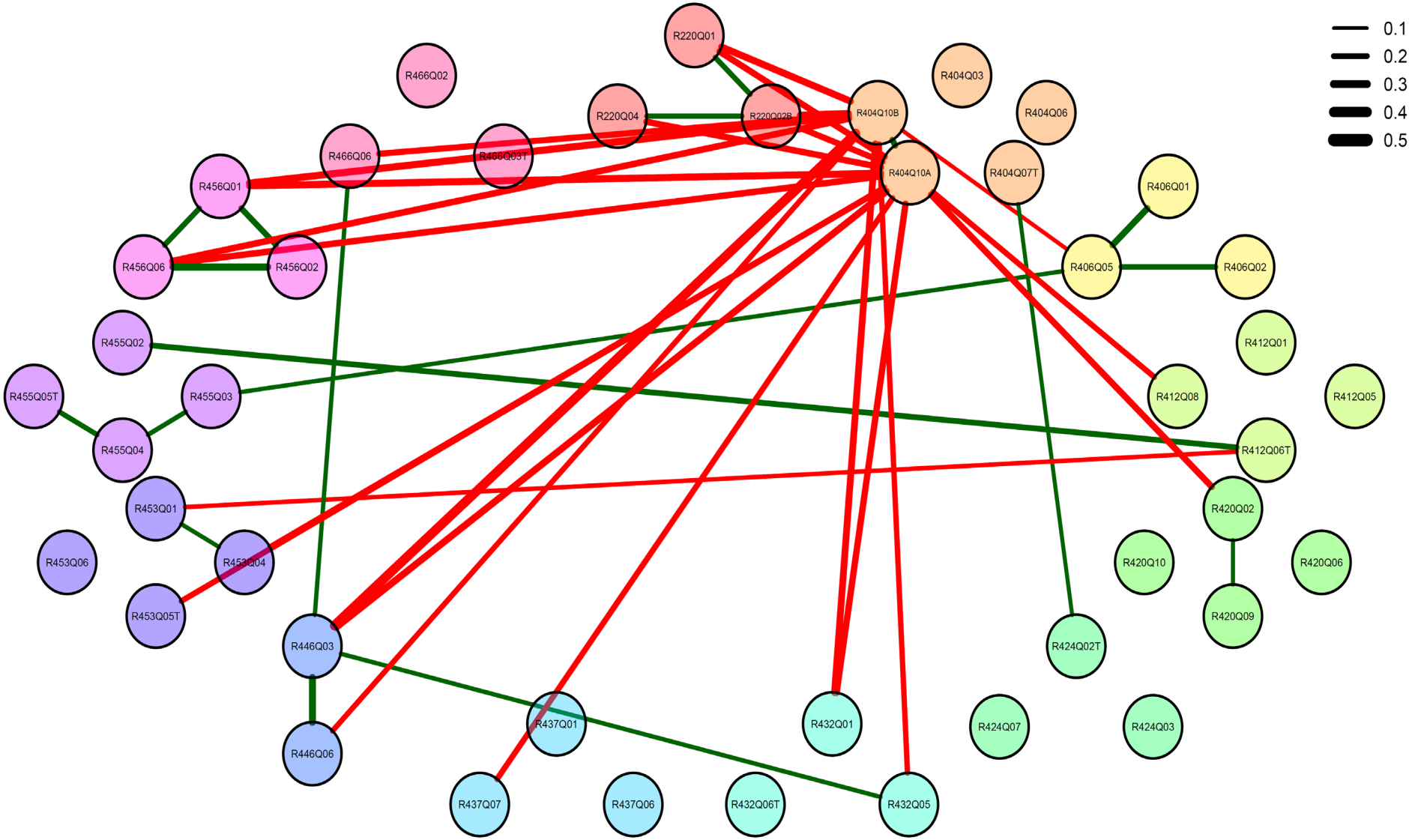


Figure 5.4.11 Visualisation of residual correlations data as a network - Reading PISA 2012

Positive LID between pairs of items within the same testlets

Thirteen pairs of within-testlet items with considerable positive LID were found (Electronic Appendix for Figure 5.4.11_POSTIVE LID_WITHIN TESTLET - PISA 2012 Reading). For ten of them, identical dependency associations were also found in the PISA 2009 study. Unfortunately, none of the testlets of interest belonged to released items groups. Seven out of 13 testlets used had at least one LID indicating a pair of items.

Positive LID between pairs of items from different testlets

Electronic Appendix for Figure 5.4.11_POSTIVE LID_BETWEEN TESTLETS - PISA 2012 Reading lists five pairs of items from non-matching testlets which produced positive LID. Given that the items cannot be previewed and the known characteristics of the questions do not offer any qualitative suggestion as to why the dependency is present, the item pairs are reported in the Electronic Appendix to facilitate a review by readers with full access to reading items.

Negative LID between pairs of items from different testlets

A pattern of negative LID can be observed for PISA 2012. There were 23 pairs of items for which RC was lower than -0.1 and MI more than 10 (see Electronic Appendix for Figure 5.4.11_NEGATIVE LID_BETWEEN TESTLETS - PISA 2012 Reading). All but one of these pairs involved either item R404Q10A or R404Q10B. None of the items involved were released. The only noticeable regularity of the information about these questions can be observed in a fact that half of the items paired with R404Q10A and R404Q10B are considerably easier, with the difference in difficulty for all 12 items exceeding 30%. Both items were also of "Constructed Response Expert"[32] item format which requires a written response from the students. Thus, it appears that pairing items of dissimilar difficulties can cause negative LID. It could be that students implement selective effort allocation, suggested by Yen (1993), and give less attention to items appearing harder in order to ensure they complete as many items as possible in the test. Alternatively, or in addition, the pairing of a 'Constructed Response Expert' item with an item using a different format may be an indication of test-wiseness of selective time allocation that in turn could give a partial

[32] In PISA 2012 the naming of the item formats changed. However "Constructed Response Expert" type of question can be seen as close to "Open constructed response" used in the first four PISA assessments.

explanation for the negative LID.

5.4.1.4 Summary and cross wave consistency of LID in the reading domain

Positive LID between pairs of items within the same testlets

Positive within-testlet dependency proved to be quite consistent across multiple PISA studies as collated in Electronic Appendix for Figures 5.4.7-11 and visualised in figures throughout Section 5.4.1.3. Used on four occasions, the three-item testlet R067 "Aesop" produced positive LID among all its questions for all instances when it was employed in the PISA assessment. Similarly, questions from R220 "South Pole" showed consistent LID in multiple PISA waves with item pair R220Q01/R220Q02B featuring five times and pair R220Q05/R220Q06 four times. Another result, consistent in four waves, originated from a pair of questions from R104 "Telephone" (R104Q01/R104Q05). Constituting three items R219 "Employment" produced positive LID on all three occasions when it was used for two pairs of its items. There was a considerable number of positive LID showing within-testlet item pairs from reading testlets, which were used only in PISA 2000 when this cognitive domain was the main focus. In this wave, there were 33 non-singular reading testlets used, and 25 of them had at least one pair of items within testlets indicating LID. For the testlet labelled R119 "Gift", five out of seven items were in LID pairs. Similarly R220 "South Pole" had four out of five of its items flagging possible common stimuli dependence. Contrastingly, other large 5 item testlets such as R088 "Labour" and R077 "Flu" flagged only one dependent pair of items. In the medium size testlets containing 3 items, some had all items indicating LID, for example, R067, while other three-item testlets such as R228 "Guide" had none. The majority of PISA 2000 results reported here concur with investigations undertaken by Monseur et al. (2011). Seventeen reading testlets, reporting at least one pair of items with LID, are highlighted by both research investigations. Publication by Kreiner and Christensen (2014), while discussing a number of limitations of PISA scaling model, also reported local dependence investigation for reading data from one booklet of PISA 2006. Results of the current study agree with a portion of the Kreiner and Christensen (2014) conclusions in regard to the existence of positive LID PISA 2006 in testlets R067, R104, R220 and even marginally for the pair from R055 (R055Q03/R055Q05) which just missed the dual-index cut point with a residual correlation of 0.099. However, Kreiner's and Christiansen's conclusions about LID in R227 "Optician" do not feature in the present study for any of the four occasions when this testlet was used. This discrepancy may be due to the fact that the Kreiner

and Christensen (2014) study used samples from 56 developed and developing countries as opposed to only 26 OECD nations utilised in this research.

As can be seen in Figure 4.3.2, the equating of PISA 2009 and 2012 was undertaken on the basis of new reading testlets introduced in 2009 with the exception of R220. Nonetheless, a cross-wave positive LID consistency was present, and can be seen in the item pairs from R406 "Kokeshi Dolls", R420 "Children's Futures", R446 "Job Vacancy", R455 "Chocolate and Health" and R456 "Biscuits". Two publications, one by Oliden and Lizaso (2013) and another by Trendtel et al. (2014) were found to discuss item dependency in the reading domain of PISA 2009. However, the studies utilised only selected national samples, Spanish and German, respectively. The first paper acknowledged the presence of dependency without listing the specific item pairs that showed it. The second publication (Trendtel et al., 2014) showed a substantial consistency with the results presented here despite different LID detection methods being used.

With regard to plausible LID drivers, the qualitative investigation offered some insight into the dependency results for PISA 2000, given that a larger proportion of testlets were released after this initial PISA assessment. These released items suggest that the need to read and refer to the introductory text, tables or figures is likely to be predominantly responsible for within-testlet positive dependency.

Positive LID between pairs of items from different testlets

Dual-index positive LID for items from different testlets was not observed in PISA 2003 and 2006 reading results. (Electronic Appendix for Figures 5.4.7-11). The remaining three PISA waves had 18 item pairs with dual-index positive LID as reported above and associated electronic appendices. Given that most of the items involved were not released to the public it is hard to provide an empirical explanation for the observed LID. However, two out of eighteen between-testlet pairs of questions came from testlets available for viewing. Both pairs featured items that are of "Open Constructed Response" type with R099Q04B present in both cases. It is speculatively proposed that dependency may be related to the scoring procedures of these open constructed response items. Issues related to scoring procedures were suggested by Yen (1993, p. 190) as one of the possible reasons for LID. There was only one case that the item pair retained positive dependency in two different PISA studies, but it involved an item pair (R406Q05/R455Q03) from non-released testlets.

Negative LID

The most striking regularity in negative dependency investigation was produced in PISA 2012 when 22 out of 23 item pairs involved either R404Q10A or R404Q10B. It was suggested that selective effort allocation, identified by Yen (1993) as a possible reason for negative LID, could be responsible for the results. At the same time, both items from R404 revealed positive LID between them so, as suggested by Habing and Roussos (2003), the need for negative LID in the presence of positive dependency also may be involved. The negative dependency in PISA 2003 and PISA 2006 showed considerable consistency involving pairs of items from R219 and R220. This may be due to the fact that both studies used precisely the same sets of reading testlets (OECD, 2009b, p. 29) allocated in the same way to two reading clusters. Both testlets were placed in cluster R1 in PISA 2003 and PISA 2006.

5.4.1.5 Qualitative investigation of reasons for LID in the science domain

As is the case when reporting mathematics and reading results above, explanations for the occurrence of LID in science is discussed separately for each wave and supported with figures and corresponding electronic appendices.

It was possible to locate only 13 released science testlets out of the 45 testlets that were used in the five waves of science assessment for the international calibration samples. This smaller number of released items is reflected in a more limited elaboration on plausible reasons for the LID observed between science items, particularly for later PISA waves. At the same time, the search for information about the items' characteristics was very successful. The PISA 2012 Technical Report (OECD, 2014b) provided some information about items that could also be propagated to earlier waves as long as PISA 2012 items were used as linking questions. However, the majority of the details about the items' characteristics (e.g. format, context, competencies) came from a Czech publication (Mandíková & Bašátková, 2008). With the help of this publication and PISAs' frameworks (OECD, 2006, 2013) as terminology references, the retrieval of the characteristics of all science items was achieved. During the writing of this section about science items, it was observed, after closer inspection of the data, that a number of item pairs only marginally missed the classification of dual-index LID. As these items frequently offered logical explanations for items dependency that involved many waves the discussion includes them and labels them as showing marginal LID. The smaller proportion of science items that were released to the public was also a

factor in investigating the marginal dependency. The limitations related to utilising the rule of thumb cut points are discussed in section 7.2.1.

PISA 2000

Figure 5.4.12 below (see also Electronic Figure 5.4.12 - Science PISA 2000) shows that in PISA 2000 there were only two pairs of science items with considerable within-testlet LID. Between-testlets, dual-index LID was identified in only a few pairs of items. Interestingly, negative LID is common and expressed mostly in between-testlet items.

Green lines – positive residual correlations
Red lines –negative residual correlations

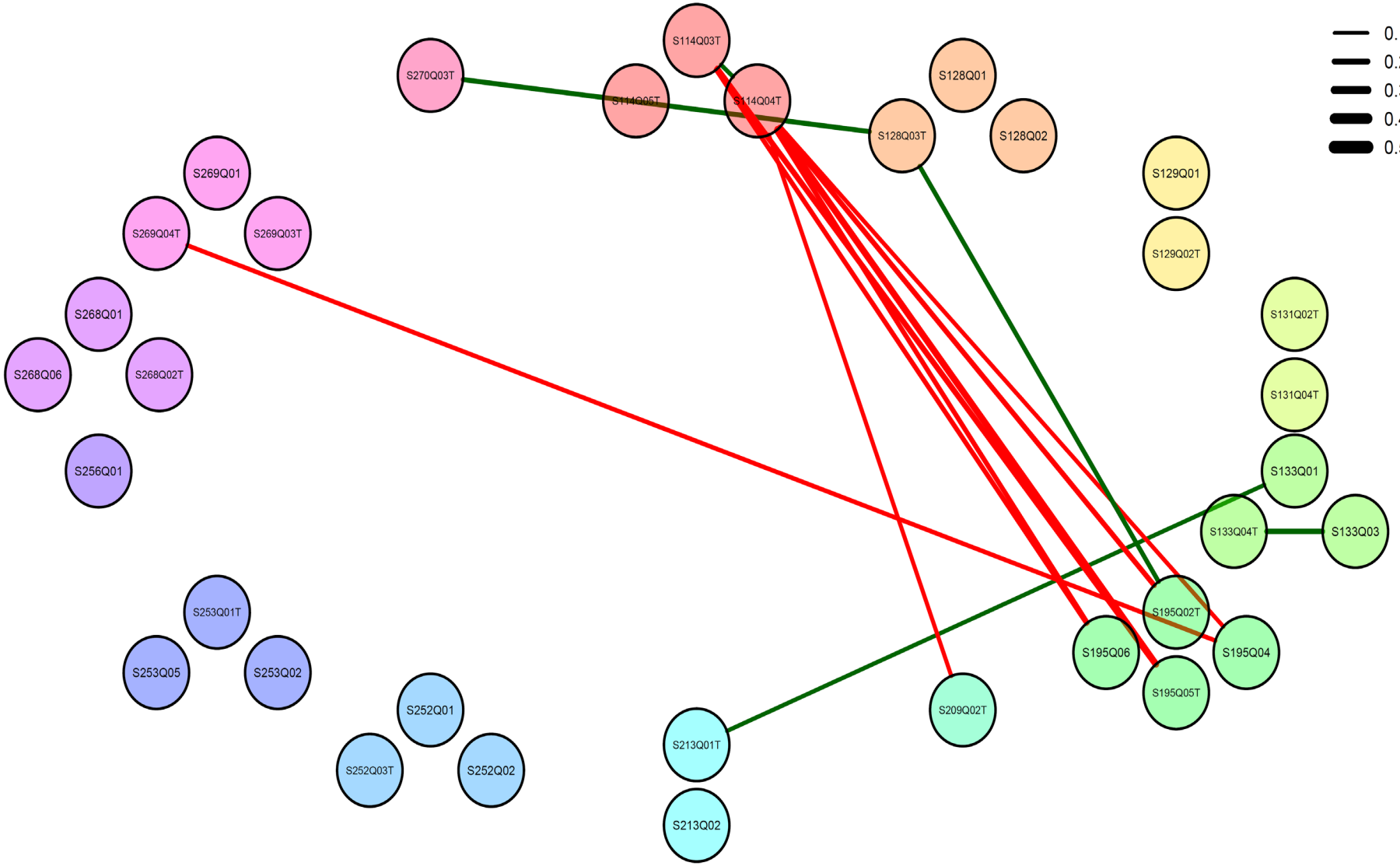


Figure 5.4.12 Visualisation of residual correlations data as a network - Science PISA 2000

Positive LID between pairs of items within the same testlets

Out of 12 non-singular testlets only two, S114 "Greenhouse" and S133 "Research", revealed considerable positive within-testlet LID (see Electronic Appendix for Figure 5.4.12_POSTIVE LID_WITHIN TESTLET - PISA 2000 Science). In the pair of released items, S114Q03T and S114Q04T, it can be seen that responding to them involves consulting a common stimulus with the trend graphs being crucial for both items. However, the answers to both items also require the same skill of graph interpretation. Moreover, both of them are open response questions. It is likely that having a common stimulus is not a single LID driver and that the skill being tested and the common response format both contribute to the observed LID. Additional investigation of the dataset of item pairs' RCs and MIs revealed five other within-testlet item pairs only marginally not meeting the requirements for dual-index LID. Given that released items are scarce for science and less likely to be available for later waves, these pairs have been additionally reported in italics in the corresponding electronic appendix to differentiate them from those pairs that do meet the dual-index LID criteria. Two items, S195Q04 and S195Q05T from the testlet "Semmelweis' Diary", are likely to produce LID mainly due to common stimuli, as without reading the preceding both of them Text2 it would not be easy to provide correct answers to both items. On the other hand, positive LID for the pair from testlet S128, "Cloning" may be more related to students' prior knowledge about cloning as the introductory reading is not required and does not explicitly give the answer for either of the questions. Another pair of items from S129 "Daylight" indicates positive LID, and the introductory reading gives no direct answers to either of the questions involved (S129Q01 S129Q02T). The likely reason for LID is the common underlying knowledge (about astronomical aspects of Earth's changing seasons, and daily time change) tested in both items.

Positive LID between pairs of items from different testlets

Three pairs of items showed considerable positive LID between items from different testlets. Once again three other pairs were close to the dual-index LID thresholds; hence all six items are reported in an electronic appendix (Electronic Appendix for Figure 5.4.12_POSTIVE LID_BETWEEN TESTLETS - PISA 2000 Science). Three out of six pairs involve items that are released. Interestingly, two pairs of these S128Q03T/S270Q03T and

S128Q03T/S213Q01T comprise items of the same format that ask participants to judge whether the statements or question are of a scientific nature. Furthermore, the third pair of known items also has one item (S128Q03T) inquiring about the scientific judgment of two listed reasons for cloning. Paired with this S128 item is S195Q02T. While S195Q02T is an open response format, it also requires students to judge and elaborate about the scientific statement that "puerperal fever is unlikely to be caused by earthquakes." A similar scientific inquiry question, S213Q01T, is also involved in another pair of items (S133Q01 and S213Q01T) which show considerable positive LID. While S133Q01 has not been released, the title of this testlet ("Research"), as well as its multiple choice item format, allows the speculation that this may be another pair of LID showing items involving questions about scientific judgement. Furthermore, both items (S133Q01 and S213Q01T) target the same item content of "Scientific enquiry" in "Knowledge about science" and both deal with the "Frontiers of science and technology" content.

It appears that for this PISA wave, an underlying reason for between-item positive LID is likely to be related to the skill of judging the process of scientific inquiry.

Negative LID between pairs of items from different testlets

Similarly in the situation in PISA 2000 mathematics, eight pairs of items from various testlets reveal negative LID (Electronic Appendix for Figure 5.4.12_NEGATIVE LID_BETWEEN TESTLETS - PISA 2000 Science). Six of them involve different combinations of pairs of items from testlets S114 "Greenhouse" and S195 "Semmelweis' Diary." Both testlets appears to require a considerable cognitive effort from students with testlet introductions requiring the reading of over 190 words and close to 290 words for S114 and S195, respectively. Furthermore, these long introductions are then followed by second prompts involving two graphs and close to 100 words of text for "Greenhouse" and over 130 words for Text 2 for "Semmelweis' Diary." Yen (1993) suggested that negative LID could be due to time management and selective effort allocation. Both testlets were presented together only to participants using PISA 2000 Booklet 8 which had a smaller number of reading clusters with a larger presence of mathematical and science items. It is possible that students who took Booklet 8 allocated a different amount of effort to these two testlets driven perhaps by their order or students' choice in preferring one testlet over the other.

PISA 2003

While science was not the target cognitive domain in either PISA 2000 or PISA 2003, some changes to the composition of the science items were made. PISA 2000 testlets S195 “Semmelweis’ Diary”, S209 “Tidal Power”, and S253/S270 “Ozone” were not included in a later wave. The new science testlets S304 “Water”, S326 “Milk”, and S327 “Tidal Energy” were utilised in PISA 2003 instead. Out of 13 science testlets used in PISA 2003, only four are available for an in-depth review. Figure 5.4.13 below and its electronic equivalent (Electronic Figure 5.4.13 - Science PISA 2003) show pairs of items in PISA 2003 that indicate dual-index LID.

Green lines – positive residual correlations
Red lines –negative residual correlations

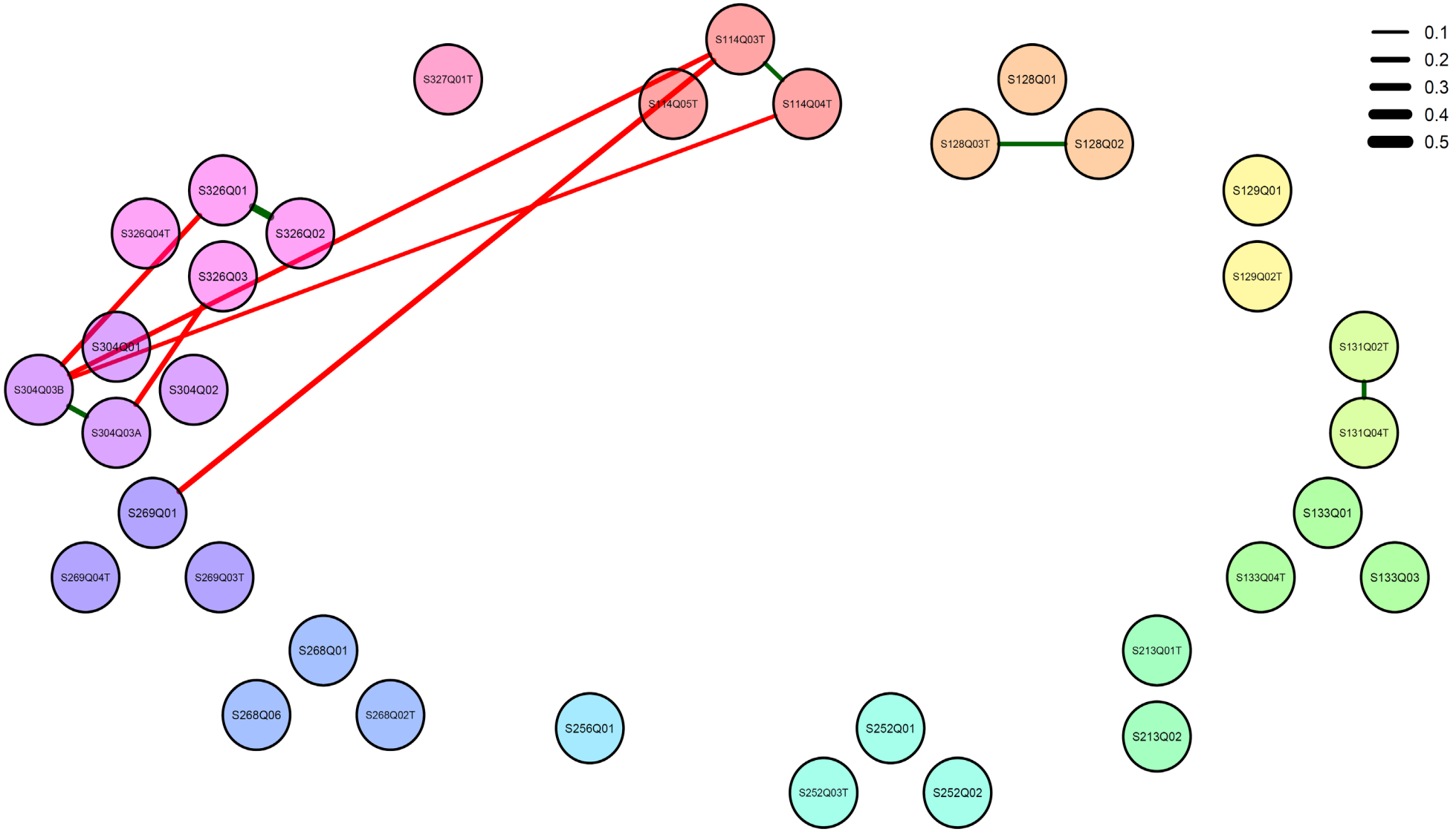


Figure 5.4.13 Visualisation of residual correlations data as a network - Science PISA 2003

Positive LID between pairs of items within the same testlets

Five pairs of items indicating dual-index LID, and a further five just missing the criteria for dual-index LID, were identified (Electronic Appendix for Figure 5.4.13_POSTIVE LID_WITHIN TESTLET - PISA 2003 Science). Two of these pairs involve the released testlets S114 "Greenhouse" and S128 "Cloning". Both pairs were also found to exhibit positive LID in a previous PISA wave. Plausible explanations for the LID are the same as reported above for PISA 2000. LID in S114 is likely due to joint stimuli of needing to refer to the graphs for both involved items. LID in S128 was argued to be more likely to arise because of common knowledge required. Three pairs of items used in both waves maintain the positive LID status. The new testlets S304 "Water" and S326 "Milk" returned the two highest RCs of 0.16 and 0.27, respectively.

Positive LID between pairs of items from different testlets

While Figure 5.4.13 does not show any between-testlet pairs of items with positive LID, five marginal dual-index LID pairs are reported in the electronic appendix (Electronic Appendix for Figure 5.4.13_POSTIVE LID_BETWEEN TESTLETS - PISA 2003 Science). Two of the five pairs (S128Q03T and S213Q01T), (S128Q03T and S268Q02T) include question S128Q03T that asks students to identify statements as being scientific and requires a Yes/No response. The same type of question is also used for S213Q01T, and this pair indicated positive LID in the 2000 wave. Thus both a common response format and a common concept could drive the observed LID. However, item S268Q02T has not been released so the explanation offered for its dependence with S128Q03T cannot be asserted. The items' formats, contexts and other item characteristics do not match.

Negative LID between pairs of items from different testlets

In PISA 2003 there are five pairs of items revealing considerable negative LID (Electronic Appendix for Figure 5.4.13_NEGATIVE LID_BETWEEN TESTLETS - PISA 2003 Science) and all of them involved items from S114 "Greenhouse", S326 "Earth's Temperature", and S304 "Water". Only items from S114 are released, and both of them are very demanding. Both require the reading of a passage with just over 200 words, followed by a second prompt involving two graphs and an additional 90 words of text, concluding with open-ended questions involving written answers that need to be justified. While no other items are released, the information about item format shows that at least one item in all five

questions' pairs was of an open response type. This may suggest that negative LID was induced by selective effort allocation leading students to be less attentive to open response questions which could appear to require more effort.

PISA 2006

Science was the targeted cognitive domain in 2006 with 103 items allocated to 36 testlets. Table 4.2.1, presented above, elaborates on which items have been used across waves for assessment linking. Only 16 items' pairs revealed dual-index LID as shown in Figure 5.4.14 (see also Electronic Figure 5.4.14 - Science PISA 2006).

Green lines – positive residual correlations
Red lines –negative residual correlations

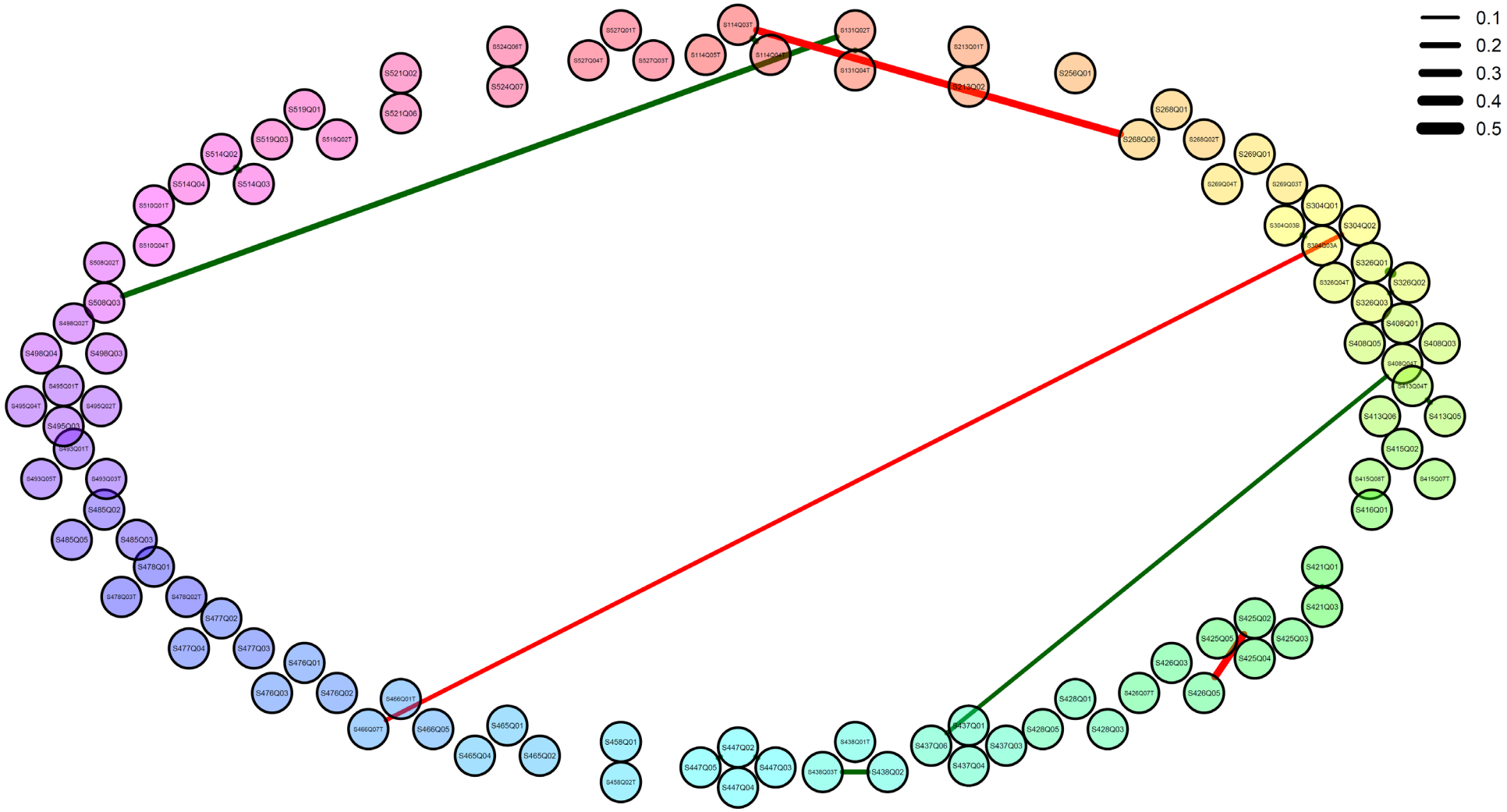


Figure 5.4.14 Visualisation of residual correlations data as a network - Science PISA 2006[33]

[33]This figure has high resolution and can be zoomed in or alternatively viewed in a corresponding electronic appendix.

Positive LID between pairs of items within the same testlets

As detailed in Figure 5.4.14 Electronic Appendix for Figure 5.4.14_POSTIVE LID_WITHIN TESTLET - PISA 2006 Science, there were eleven couplings of items from the same testlets making the dual-index LID cut point. However, just as many pairs were very close to making the cut with their RCs and MIs in excess of 0.8 and 8 for RC and MI, respectively. Consequently, the electronic appendix mentioned above shows the details for 22 pairs of items. Most of the released science items were introduced in PISA 2000, and there are very few available for viewing from later waves of the study. Details of only five item pairs originating from only two testlets are available. Positive LID for S114Q03T and S114Q04T from "Greenhouse" testlet is once again present, replicating the PISA 2000 and 2003 conclusions. The only other testlet which could be closely scrutinised is S447 "Sunscreens", which was introduced in PISA 2006. Three combinations[34] of its items are among 22 pairs of listed pairs of items. After reading the testlet prompt and questions, arguably all the questions reference the prompt. A pair of items, S326Q01 and S326Q02 from the "Milk" testlet, are discussed in the positive within-testlet section for PISA 2003. Again, they produced the highest RC out of all 22 positive LID. Due to the confidential nature of this testlet, the reasons for positive LID for this pair of items cannot be elaborated upon. Other pairs, (S304Q03A/S304Q03B) and (S131Q02T/S131Q04T), reported in PISA 2003, once again reveal dual-index positive LID. Furthermore, two other pairs, (S304Q01/S304Q02) and (S269Q01/S269Q03T), showed marginal dual-index LID in a prior wave and retained their status in PISA 2006. Positive within-testlet LID was also present for testlets introduced in PISA 2006, in particular, the pairs of items from S514 "Development and Disaster", S438 "Green Parks", S413 "Plastic Age" and S421 "Big and Small." Because they have not been released, no basis for inferring possible drivers for the observed LID can be offered at this stage.

Positive LID between pairs of items from different testlets

Two pairs of items produced RCs and MIs exceeding 0.1 and 10, respectively, and an

[34] The testlet S447 "Sunscreens" included four items in PISA 2006 labelled S447Q02, S447Q03, S447Q04, S447Q05T. The only located source (OECD, 2009d) from which the information about these released items could be extracted does not use the official labels and only numbers them according to their order. It is believed that matching in the electronic appendix Electronic Appendix for Figure 5.4.14_POSTIVE LID_WITHIN TESTLET - PISA 2006 Science is correct more so that the known items' formats can be used to verify the items allocation.

additional nine were very close to these thresholds. All of these, along with item characteristics, are reported in Electronic Appendix for Figure 5.4.14_POSTIVE LID_BETWEEN TESTLETS - PISA 2006 Science. None of the pairs involves both items from released testlets limiting any investigation of plausible reasons for their dependency. However, in the case of pair S426Q07T and S495Q04T, for which one question from S495 "The Grand Canyon" is released, informed speculation can be undertaken. Item S426Q07T is of YES/NO type, requiring students to judge the statements in regard to their scientific nature. This question item type is "Complex Multiple Choice", its item content dealing with "Knowledge about science - Scientific enquiry", and its category of scientific cognitive processes places this item into "Identifying scientific issues" scientific competency (OECD, 2006, p. 29). Interestingly, the similar YES/NO scientific judgement format released questions (S128Q03T, S213Q01T, S270Q03T) discussed above in two "*Positive LID between pairs of items from different testlets*" sections all featured the same three item dimensions as S426Q07T. By matching item properties, it is likely that S495Q04T from S495 "Radiotherapy" is of the same YES/NO judgemental type. Furthermore, S495Q04T features in another positive LID pair with item S415Q07T from "Solar Panels" testlet and once again the matching item characteristics are observed. It may be noted that another pair S438Q01T from "Green Parks" and S466Q07T from "Forest Fires" once again classify both items as "Complex Multiple Choice", "Knowledge about science - Scientific enquiry" and "Identifying scientific issues". The ability to judge scientific statements once again appears to relate to between-testlets positive LID. The value of detailed qualitative investigation proves to be particularly evident in this case as different sources of evidence for dependence as well as the cross-wave consistency leads to the strengthened conclusion that skills in "Identifying scientific issues" competency drives many of LID cases linking items from different testlets.

Negative LID between pairs of items from different testlets

Only three pairs of PISA 2006 scientific literacy items showed considerable negative LID as reported in the Electronic Appendix for Figure 5.4.14_NEGATIVE LID_BETWEEN TESTLETS - PISA 2006 Science. None of the three pairs has both items released. None of the item characteristcs of any of the six items suggested plausible explanations for negative LID. However, item S114Q03T featured as being involved in negative LID in PISA 2000 and PISA 2003 as reported above.

PISA 2009

Science in PISA 2009 was not a main investigated literacy, and all items used in this iteration of the PISA were also used in 2006 assessment. Figure 5.4.15 (see also Electronic Figure 5.4.15 - Science PISA 2009) shows only four pairs of items constituting considerable positive LID within the same testlets, one positive LID item pair linking items from different testlets and six negative LID couplings of science questions.

Green lines – positive residual correlations
Red lines –negative residual correlations

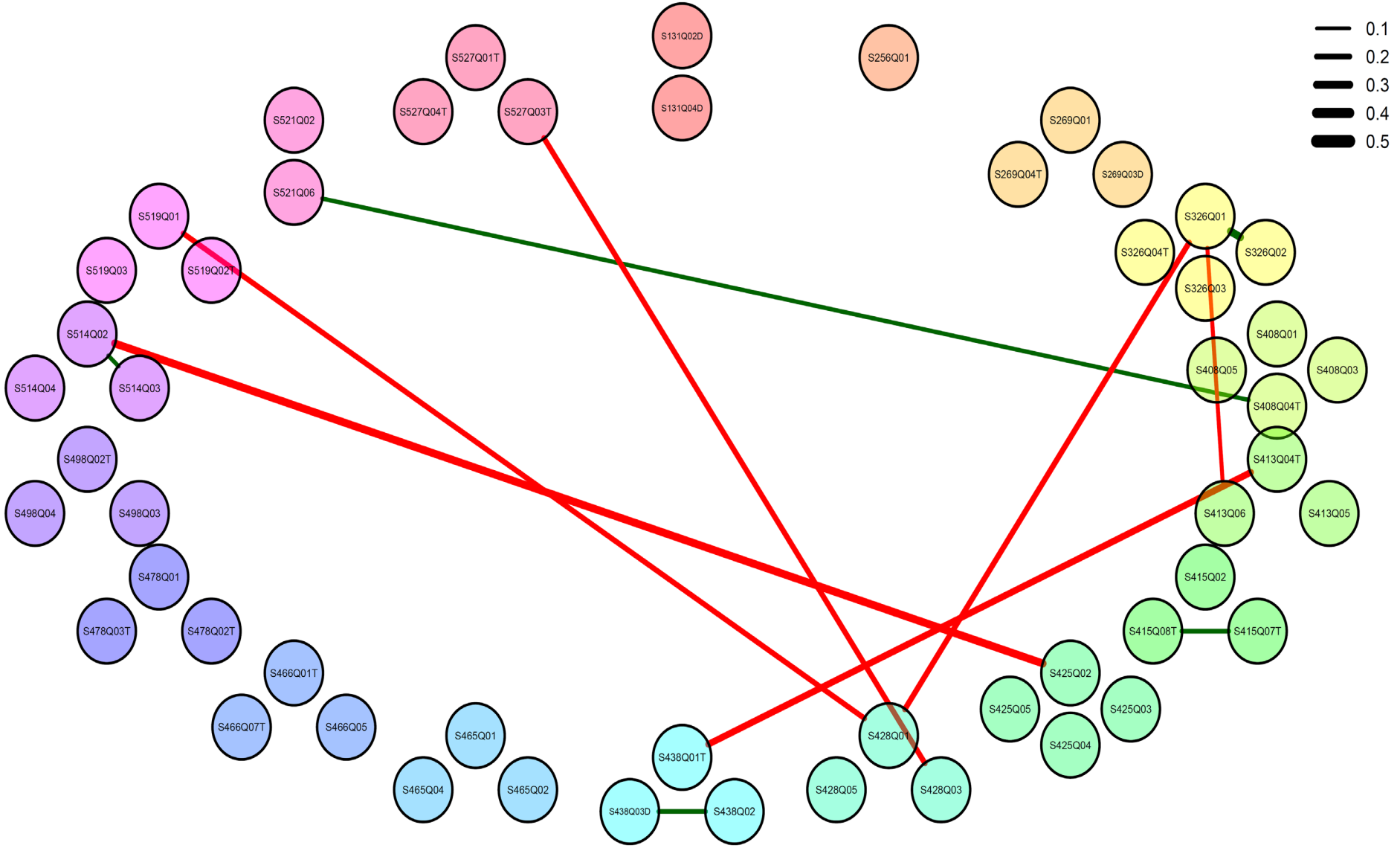


Figure 5.4.15 Visualisation of residual correlations data as a network - Science PISA 2009

Positive LID between pairs of items within the same testlets

Four item pairs were found indicating dual-index LID (Electronic Appendix for Figure 5.4.15_POSTIVE LID_WITHIN TESTLET - PISA 2009 Science). As in PISA 2003 and PISA 2006, pair S326Q01 and S326Q02 from the "Milk" testlet once again produced the largest RC in PISA 2009. In the same manner pairs S415Q07T/S415Q08T from "Solar Power Generation," S438Q02/S438Q03 from "Green Parks" and S514Q02/S514Q03 from "Development and Disaster" featured in PISA 2006. An additional five positive LID item pairs listed in the featured electronic appendix only minimally missed the dual-index LID cut-point. All five of them were reported at least once in the previous three waves as revealing positive LID among within-testlet items. As none of the pairs discussed here was released, no plausible explanation for underlying positive LID can be offered.

Positive LID between pairs of items from different testlets

Electronic Appendix for Figure 5.4.15_POSTIVE LID_BETWEEN TESTLETS - PISA 2009 Science lists only one pair of items, S408Q04T from S408 "Wild Oat Grass" and S521Q06 from S521 from "Cooking Outdoors" as displaying dual-index LID. However, ten other pairs are also reported in the electonic appendix. None of 22 involved items were released. Looking at the items' characteristics for items' pairs also does not reveal any obvious regularities.

Negative LID between pairs of items from different testlets

Six pairs of items across different testlets produced negative LID (Electronic Appendix for Figure 5.4.15_NEGATIVE LID_BETWEEN TESTLETS - PISA 2009 Science) with the largest RC value being -0.28 for pair S425Q02 and S514Q02 from testlets "Penguin Island" and "Development and Disaster", respectively. None of the items involved was released, and the only notable characteristic for these two items is that S514Q02 was very simple to answer with 84% students providing a correct answer while the percentage of students giving a spot-on response to the S425Q02 was slightly below 50%.

PISA 2012

The collection of science items and make-up of testlets for PISA 2012 were identical to those in PISA 2009. While some LID indicating pairs of items reproduce the previous wave results, others do not, and that is particularly visible in the case of negative LID. Figure

5.4.16 (see also Electronic Figure 5.4.16 - Science PISA 2012) reports PISA 2012 dual-index LID in more detail. It may be somewhat surprising that the dependency reporting graphs for PISA 2009 and 2012 are considerably different given the matching sets of science items. However, while the PISA 2012 Technical Manual (OECD, 2014b) acknowledges that three science clusters were not changed[35] and cluster rotation design for the standard booklets corresponds to this in PISA 2009, a closer inspection of the technical manuals for both studies reveals a change that could be important for the item dependency. Figure 2.1 in OECD (2014b, p. 31) and Table 2.1 in OECD (2012, p. 30) show the allocation of various clusters to booklets. In PISA 2009 all three booklets (Bk3, Bk10, Bk12) with two science clusters are accompanied by one reading and one mathematics cluster. This is also the case in PISA 2012 for Booklet 8 and Booklet 12, but not for Booklet 1. PISA 2012 Booklet 1 has two clusters with science items and two clusters with mathematics items. This may cause a larger quantitative burden for some students, and it also highlights how these two instances of science evaluation differ despite matching sets of items. Subsequent sections report in detail about the three types of LID allocation, as is done in the corresponding sections above.

[35] The names of the clusters changed, but the testlets allocation within the science clusters remained the same.

Green lines – positive residual correlations
Red lines –negative residual correlations

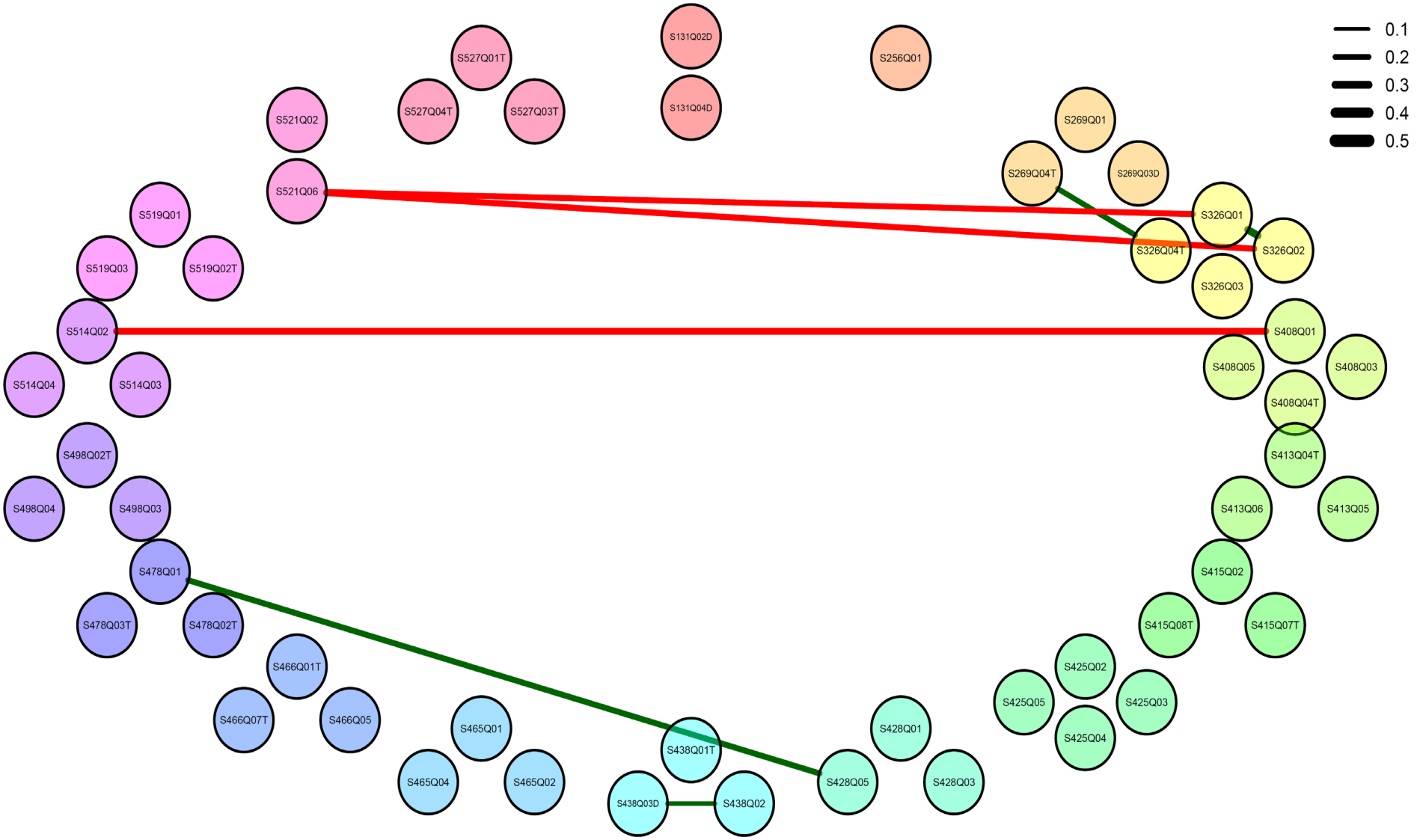


Figure 5.4.16 Visualisation of residual correlations data as a network - Science PISA 2012

Positive LID between pairs of items within the same testlets

Once again, as in all the datasets since PISA 2003, pair S326Q01 and S326Q02 of two items from "Milk" produced RCs close to 0.3 (see Electronic Appendix for Figure 5.4.16_POSTIVE LID_WITHIN TESTLET - PISA 2012 Science). Also, pair S438Q02 and S438Q03 from "Green Parks" indicate considerable positive LID. Items S514Q02 and S514Q03 from "Development and Disaster" which were featured since its introduction this time are only slightly short of making the dual-index LID category. A similar pair from the "Earth's Temperature" testlet (S269Q03/S269Q04) on this occasion is borderline positive LID, as it was in PISA 2000 and PISA 2009. Other 'three item' pairs are also reported in the electronic appendix as revealing LID, but none of them are available for inspection. Qualitative analysis and an in-depth review of the data offered here supports an arguement for cross-wave consistency in science items dependency.

Positive LID between pairs of items from different testlets

Two pairs of items from different testlets were identified as showing considerable positive LID (S428Q05/S478Q01 and S269Q04/S326Q04). The former pair was also reported in the previous wave. An additional four pairs with marginally positive LID were not identified in the corresponding section above (see also Electronic Appendix for Figure 5.4.16_POSTIVE LID_BETWEEN TESTLETS - PISA 2012 Science). As was the case for mathematics and reading evaluations, the later waves' qualitative investigations involved far fewer released items, limiting the scope for qualitative investigations into the drivers of dependency. However, the Electronic Appendices were also includedfor the facilitation of further investigations by readers with access to restricted items and testlets.

Negative LID between pairs of items from different testlets

The Electronic Appendix for Figure 5.4.16_NEGATIVE LID_BETWEEN TESTLETS - PISA 2012 Science reports three pairs of items indicating negative LID, none of which are released. However, it is known that in all three pairs, one of the items is of the Open Response type offering the possibility of selective effort allocation, which was suggested previously as a plausible negative dependency trigger.

5.4.1.6 Summary and cross wave consistency of LID in the science domain

Positive LID between pairs of items within the same testlets

Paralleling the previous two summaries for mathematics and reading, the qualitative science conclusions will focus on cross-wave consistency. What differentiates this cognitive domain is the introduction of marginal dual-index LID, driven by a visual inspection of residual correlations science data as well as the scarcity of released items. The electronic appendix (see Electronic Appendix for Figures 5.4.12-16) reproduces all the results permitting more efficient inspection of dependency across multiple PISA waves. The pair of items (S326Q01 and S326Q02) from the non-released four-item testlet S326 "Milk", proved to produce the largest residual correlations of about 0.3 consistently in all four PISA studies which used this testlet. High cross-wave consistency was also shown in case of testlet item pair S438Q02 and S438Q03, belonging to the "Green Parks" and used in PISA 2006, 2009 and 2012. The third pair of positive LID items used for cross-wave linking came from S114 "Greenhouse", which was used on three occasions. It was argued for these released items that the dependency is likely to be due to a common graph required for both questions. Alternatively, some positive LID in released items appeared to be unrelated to the testlet introduction as was the case for item pair (S128Q02 and S128Q03) from S128 "Cloning". It was argued that underlying knowledge about the topic might be a more likely dependency driver. The largest number of within-testlet pairs of items with positive LID was evident in PISA 2006 when science was the main tested domain. There is only one publication that could be used as confirmation of these results (Le Hebel, Montpied, Tiberghien, & Fontanieu, 2017). Although dependency on testlet stimulus is not of primary interest in this paper by Le Hebel et al. (2017), the authors had access to all science items from PISA 2006 and reported on a 'three points categorical scale' for their perceived level of item dependence on the information in the testlet introduction. Out of eleven pairs of PISA 2006 items presenting positive within-testlet LID, seven involved both items labelled by Le Hebel et al. (2017) as dependent on the testlet stimulus. This does not disregard the remaining four items' doublets as their LID may be driven by other reasons, for example, item chaining.

Positive LID between pairs of items from different testlets

The existence of positive LID between items from different science testlets proved to be of interest, due to obtaining access to a few key items as well as utilising item characteristics for non-released testlets. Originating in PISA 2000, two pairs of items: S128Q03T/S270Q03T and S128Q03T/S195Q02T came from openly available testlets. It was argued that their dependency was driven by a requirement to arbitrate whether the listed statements are of a scientific nature following a Yes/No question format. PISA 2003 also confirmed this regularity when a similar pair of questions, S128Q03T and S213Q01T, featured as indicative of positive dependency. Results from PISA 2006 add to this consistency. Although the four questions listed above were not used in this

wave, S426Q07T, an item new to PISA 2006, was also of the same type. It was discovered that all the above mentioned and accessible five scientific judgement question were of the same item type, i.e. “Complex Multiple Choice”, item content, i.e. “Knowledge about science - Scientific enquiry” as well as the category of scientific competency, i.e. “Identifying scientific issues” (OECD, 2006, p. 29). Through the deduction and item properties matching, it was suggested that features such as between-testlet LID items, S495Q04T from S495 “Radiotherapy”, S415Q07T from “Solar Panels”, S438Q01T from “Green Parks” and S466Q07T from “Forest Fires”, may all be Yes/No judgement type. It is very likely that this kind of question and more general scientific competency of “Identifying scientific issues” create positive LID. Cross-wave consistency (see also Electronic Appendix for Figures 5.4.12-16) is limited to two pairs featured in two PISA waves, namely S128Q03/S213Q01, discussed above, and S428Q05/S478Q01. The last item pair comes from non-released testlets S428 “Bacteria in Milk” and S478 “Antibiotics”. However, the titles of the testlets suggest that positive dependency may relate to knowledge shared by these two testlets.

Negative LID

No cross-wave consistency is apparent for negative dual-index LID, as can be seen in the relevant paragraphs of Section 5.4.1.5 which are also available in an aggregated form in Electronic Appendix for Figures 5.4.12-16). However, it appears that the pairs of items presenting negative LID repeatedly involve at least one item which produced a positive within-testlet dependency. This regularity is particularly visible in PISA 2000 with S114Q03 and S114Q04 items, but also in later waves. The most likely explanation of the majority of negative LID pairs in science points to being a mathematical artefact of positive LID as suggested in the literature (Habing & Roussos, 2003; van Rijn & Rijmen, 2015). The discussion on drivers of science negative LID continues in the quantitative section of this chapter.

5.4.2 Quantitative investigation of LID drivers based on various PISA item characteristics

The purpose of this section is to extend the discussion regarding the plausible drivers of positive and negative LID and to address the previous section’s (see subchapter 5.4.1) limitations concerning the restricted number of testlets that were available for qualitative investigation. Random intercept multilevel logistic regression was utilised for the quantitative investigation, as described in detail in section 3.3.1. This section is organised into parts referring to two binary outcome measures. The first section reports three models investigating three cognitive domains predicting positive dependency, i.e. residual correlations exceeding +0.1. The second half of this section also presents a model for mathematics, reading and science but on this occasion predicting

negative dependency, i.e. residual correlations below -0.1 threshold. The choice of the LID index is elaborated in sections 3.3.3.1 and 3.3.3.2, while arguments against multinomial multilevel logistic regression are put forward in section 3.3.3.4.

The models were run hierarchically and are organised in sections which are reported in full in the associated electronic appendices. Full details showing the evolution of the models are presented in electronic appendices rather than in the body of the thesis because of the large volume of material. However, the final models are presented in the text below.

Section A of the electronic appendices reports on analyses using variables that are common across all models for the three cognitive domains. The variables included in this part look at item pairs in regard to the same submission source, submission language background, size of the testlet[36] from which item pairs originated or items being coded as binary or polytomous. A variable describing combinations of item formats was also prepared with the base category representing both items being "Simple multiple choice". Due to inconsistencies in item format terminology across different PISA waves, logical aggregations were used, for example, treating four differently labelled types of "Short Response", "Closed Constructed Response", "Constructed Response Auto-coded", "Constructed Response Manual" are all categorised as being "Short response". The limitations arising from this approach are discussed in section 7.2.1. The final variable used consistently in all six models and reported in this sub-chapter aims to quantify item pairs in regard to average difficulty as well as difficulty discrepancy of the item pair. Figure 5.4.17 visualises how six categories of this variable were derived, allocating items pairs into groups based on their average difficulty and difficulty difference between them.

[36] This variable has three levels (i.e. Both items from small testlets (base), Both items from large testlets, One item from small with second item from large testlet). However, the allocation to small versus large was different from domain to domain due to different medians for the sizes of the testlets. In mathematics testlets with 3 or more items were treated as large in this domain while for reading and science the corresponding threshold value for large testlet was 4.

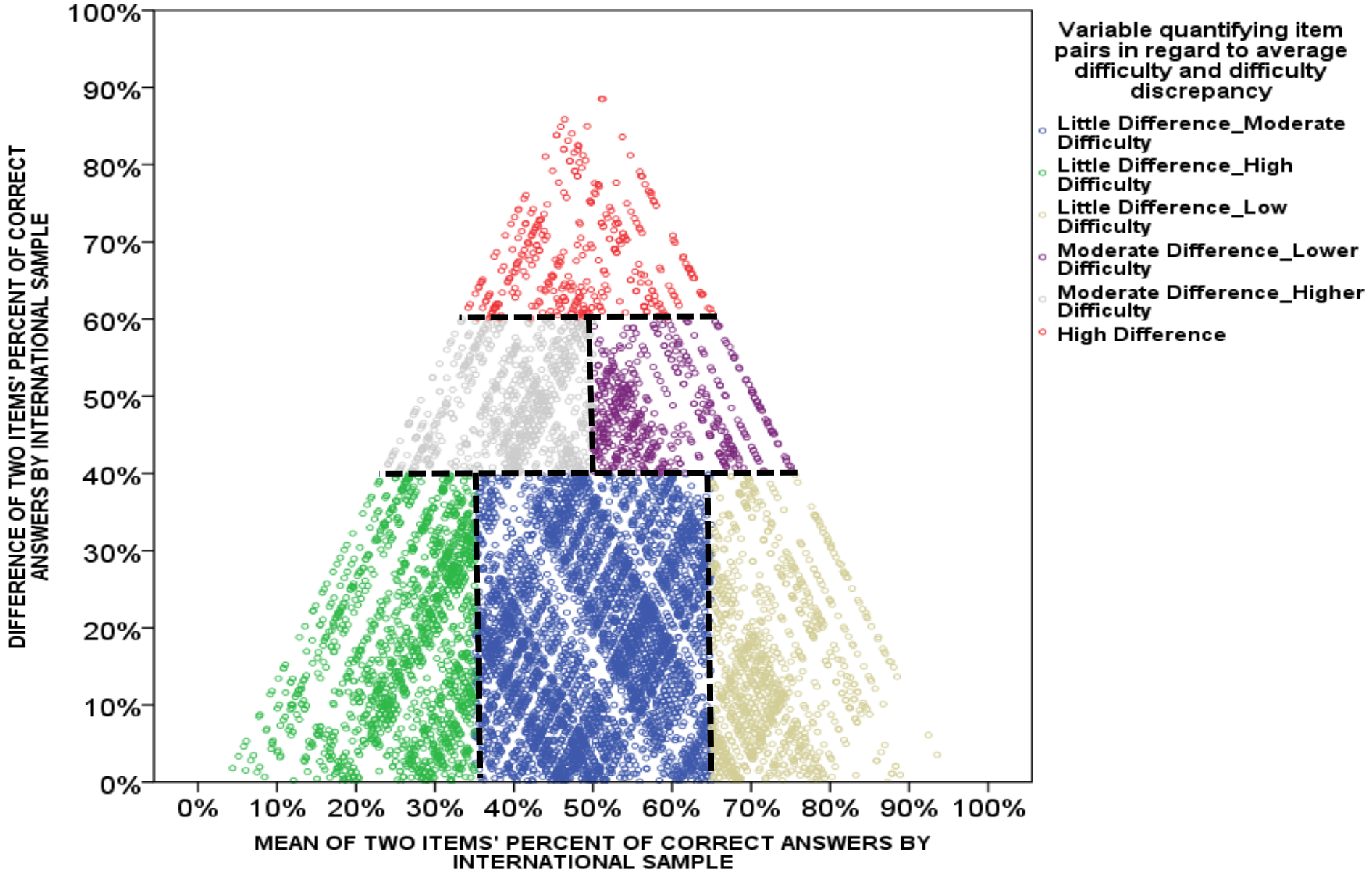


Figure 5.4.17 Visual presentation of categories for a variable used in modelling item pairs average difficulty and difficulty difference

Section B of electronic appendices reports on a series of models utilising variables from section A, with one variable being added at the time. Models in section C mimic the model construction of section A but use variables reflecting the properties of items that are unique to each cognitive domain. An annotated list of all explanatory variables specific to each cognitive domain is included in the models for mathematics, reading and science. Section D reports models with all section A variables with one of the domain specific variables being used at the time. Section E builds hierarchical models to incorporate all the variables that are available. Finally, AIC and BIC information criteria are used to propose final models which are reproduced in the text.

5.4.2.1 Models explaining positive LID

Mathematics

Beyond variables common to all cognitive domains, item characteristics specific to mathematics were used in multilevel logistic regressions to model positive LID in mathematical data. One of the specific predictors is a variable investigating pairs of items identified by their length. This variable was based largely on information about the items from an OECD publication (OECD, 2010b). Another variable describing item pairs took advantage of information about the competency being assessed by the item, and this was matched to PISA 2000's three level classification constituting *Reproduction, Connection and Reflection*. It is acknowledged that obtaining cross-wave consistency in mathematical competency may be controversial due to a non-matching approach in PISA frameworks (Niss, 2015). Cross-wave matching was undertaken based on OECD (2004, p. 49) and by means of linking item competency classifications. Similar challenges due to the evolution of PISA's mathematical frameworks (Stacey & Turner, 2015b) were faced in devising variables consistent across five waves, involving mathematical processes *(Formulate, Employ and Interpret),* mathematical content (*Quantity, Change and relationships, Space and shape, Uncertainty and data*), mathematical branch (*Number, Statistics/Probability/Data, Measurement, Algebra/Functions/Discrete Mathematics, Geometry*) or context i.e. situation (*Personal, Public, Scientific, Occupational and Educational*).

Electronic Appendix for Table 5.4.1 - Mathematics - Positive LID reports a total of 43 models with the final one being reported in the Table 5.4.1. This electronic appendix also presents[37] melogit estimated confidence intervals for each model. Null model section criteria of Akaike information criterion (AIC)=2683 and Bayesian information criterion (BIC)=2697 while corresponding results for the final model were AIC=2346 and BIC=2606.

[37] To view CIs for all models unhide columns option in MS Excel will need to be applied first.

Table 5.4.1 Final multilevel logistic regression model predicting positive LID in mathematics

		Model FINAL	95% CIs
PISA WAVE	PISA WAVE=2000 (base)	1.0	
	PISA WAVE=2003	1.0	[0.4,2.5]
	PISA WAVE=2006	1.2	[0.4,3.1]
	PISA WAVE=2009	0.6	[0.2,1.8]
	PISA WAVE=2012	0.7	[0.3,1.9]
LOCATION OF ITEM PAIRS	Items are not in the same cluster (base)	1.0	
	Items share the cluster but not testlet	0.1***	[0.0,0.2]
	Items are in the same testlet	182.3***	[55.5,598.5]
VARIABLE QUANTIFYING ITEM PAIRS IN REGARD TO AVERAGE DIFFICULTY AND DIFFICULTY DISCREPANCY	Little Difference_Moderate Difficulty (base)	1.0	
	Little Difference_High Difficulty	5.3***	[2.9,9.7]
	Little Difference_Low Difficulty	5.1***	[2.9,8.7]
	Moderate Difference_Lower Difficulty	2.5*	[1.2,5.2]
	Moderate Difference_Higher Difficulty	3.5***	[1.7,7.2]
	High Difference	15.8***	[6.6,37.8]
SIZE OF THE TESTLET	Both items from small testlets (base)	1.0	
	Both items from large testlets	0.3***	[0.2,0.6]
	One item from small with second item from large testlet	0.6*	[0.4,1.0]
ITEM FORMAT	Both items "Simple Multiple Choice" (base)	1.0	
	Both items ("Complex Multiple Choice" or "Short Response")	3.1*	[1.2,7.5]
	Both items "Open Constructed Response"	1.4	[0.4,4.7]
	Pair of "Open Constructed Response" and "Simple Multiple Choice" items	2.4	[0.9,6.4]
	Pair of "Simple Multiple Choice" and ("Complex Multiple Choice" or "Short Response") items	1.2	[0.5,3.0]
	Pair of "Open Constructed Response" and ("Complex Multiple Choice" or "Short Response") items	0.7	[0.3,1.9]

TYPE OF ITEM PAIRS	**Both Items Binary (base)**	**1.0**	
	One Item Binary One Polytomous	**0.4*** **	[0.2,0.6]
	Both Items Polytomous	**1.1**	[0.2,5.7]
"BRANCH"/STRAND OF MATHEMATICS	**Both items Number (base)**	**1.0**	
	Both items Algebra/Functions/Discrete Mathematics	**2.2**	[0.5,10.0]
	Both items Geometry	**4.1***	[1.3,12.9]
	Strand for at least one item is missing	**2.5***	[1.0,5.9]
	Both items Statistics/Probability/Data	**5.0*** **	[2.0,12.7]
	Both items Measurement	**2.5**	[0.3,23.8]
	Pair of Algebra/Functions/Discrete Mathematics and Geometry items	**1.0**	[0.3,3.4]
	Pair of Algebra/Functions/Discrete Mathematics and Number items	**0.9**	[0.3,2.6]
	Pair of Algebra/Functions/Discrete Mathematics and Statistics/Probability/Data items	**1.2**	[0.4,3.6]
	Pair of Algebra/Functions/Discrete Mathematics and Measurement items	**0.3**	[0.0,1.9]
	Pair of Geometry and Number items	**0.8**	[0.3,1.9]
	Pair of Geometry and Statistics/Probability/Data items	**0.6**	[0.2,1.8]
	Pair of Geometry and Measurement items	**0.8**	[0.2,3.5]
	Pair of Number and Statistics/Probability/Data items	**1.9**	[0.9,4.2]
	Pair of Number and Measurement items	**0.4**	[0.1,1.5]
	Pair of Statistics/Probability/Data and Measurement items	**0.2**	[0.0,1.1]

** p<0.05, ** p<0.01 *** p<0.001 | Number of item pairs=8385*

As can be seen in Table 5.4.1, none of the PISA waves (i.e. 2003, 2006, 2009 and 2012) produces significant odds of positive LID as compared to the first study in 2000. This has been consistent across all the models leading to the final one.

Consistently, but not surprisingly, the odds of positive LID were very high when pairs of items from the same testlet are compared to the reference category representing pairs of items from different clusters. The odds ratio for the final model is 182 (56, 599).

A categorical variable representing items' relative difficulty of the paired items and location gave interesting results. After controlling for other variables, the odds ratio for pairs of items that differed substantially from each other against pairs with little difference and moderate difficulty resulted in OR=15.8 (6.6, 37.8). Bonferroni-corrected pairwise comparisons[38] also show that odds of finding positive LID in item pairs that were further apart (High Difference) were higher 6.3 (1.4, 28.4) against "Moderate Difference_Lower Difficulty" as well as against "Moderate Difference_Higher Difficulty" with odds equal to 4.5 (1.1, 17.9), while controlling for other variables in the final model. It is possible that testlets were designed so the common stimuli effect was more pronounced for questions further apart in difficulty. However, as suggested by Yen (1993), external assistance or pre-knowledge would most likely emerge for items that greatly differ in their levels of complexity.

The final model also suggests that the odds of finding positive LID are smaller than one when both items are from larger testlets compared to both items originating from small testlets. Qualitative investigations reported in section 5.4.1.1 showed quite a few two item-testlets producing positive LID. Perhaps small testlets were designed in this way so the questions can "get value" out of the introductory prompt, being that larger testlets are less likely to use a common prompt.

The variable describing combinations of item formats showed OR of 3.1 (1.2, 7.5) when comparing items which were both "Complex Multiple Choice" or "Short Answer" against both items being "Simple Multiple Choice". Item or response format was identified as a possible LID generator in the Yen (1993) pivotal paper. The qualitative investigation in section 5.4.1.1 suggested, on the basis of a limited number of released items, that in some cases the shared testlet introduction has little relevance to pairing items, compared to requiring a specific mathematical skill. This occurred, for example, for items requiring converting equations which are more likely to be of "Short Answer" format.

Another variable which was included in the final model codes binary versus the polytomous

[38] Not reported but available upon request

question types. The model suggests that the odds of finding positive LID are reduced 0.4 (0.2 0.6) after controlling for other variables for differently coded items in reference to both items being binary. The polytomous items are typically open-ended questions which have to be manually scored by the raters. Yen (1993) points to the possibility that LID can be produced due to common scoring rubrics. However, it is hard to imagine how this LID driver would explain this result. The last variable included in the final model elaborates on the mathematical branch that the pairs of items are coming from. The odds of finding positive LID were higher 4.1 (1.3, 12.9) and 5 (2, 12.7) when both items are from the "Geometry" or "Statistics, Probability and Data" strand, respectively, compared to a pair of "Number" type of the cognitive questions. This result is somewhat supported by the results presented in section 5.4.1.1 that reviewed the reasons for LID in pairs of released items. Positive LID between items involving the imagining of 3D objects in 2D, as well as items requiring the ability of mean calculation, was clearly featured.

Reading

Three additional characteristics distinctive to reading were used in multilevel models for this cognitive domain. One of these was a categorical variable quantifying the situational context of item pairs, and it used four types of text situational placements *(Personal, Public, Occupational and Educational)* as introduced in the first PISA wave (OECD, 1999, p. 13) and remained unchanged for later waves (OECD, 2010a). Reading text-type was also used in the logistic regressions with three distinct types (*Non-Continuous Text (Map, Chart, Table or Form), Continuous Text that is Narrative or Descriptive or Instruction and Continuous Text that is Expository or Argumentative*). This variable was aggregated from a larger number of text types proposed in the original reading literacy framework (OECD, 1999, p. 27). Furthermore, when reading was for the second time a main PISA targeted literacy, additional changes to the types of text were made by allowing for non-continuous texts to also have different purposes (i.e. descriptive, argumentative, instructional) (OECD, 2010a, p. 32). The text types from PISA 2009 and 2012 were coded as a three level categorical text-type variable to maintain cross wave consistency. Lastly, a specialised reading variable used in the model involved three reading aspects (i.e. *Access and retrieve, Integrate and interpret and Reflect and evaluate*) as per the PISA 2009 framework (OECD, 2010a, p. 32).

Following prior analyses, template 31 multilevel hierarchical logistic regressions are reported in the Electronic Appendix for Table 5.4.2 - Reading - Positive LID with the last two columns presenting estimates for the proposed final model along with its CIs. The results of the final model are also presented below in Table 5.4.2. Final model information criteria were AIC=2771 and BIC=3035. In comparison, corresponding null model statistics were AIC=3189 and BIC=3204.

Table 5.4.2 Final multilevel logistic regression model predicting positive LID in reading

		Model FINAL	95% Cis
PISA WAVE	PISA WAVE=2000 (base)	1.0	
	PISA WAVE=2003	0.4	[0.1,1.2]
	PISA WAVE=2006	0.8	[0.3,2.0]
	PISA WAVE=2009	2.1***	[1.4,3.1]
	PISA WAVE=2012	1.3	[0.7,2.3]
LOCATION OF ITEM PAIRS	Items are not in the same cluster (base)	1.0	
	Items share the cluster but not testlet	0.2***	[0.1,0.5]
	Items are in the same testlet	57.5***	[24.1,137.6]
VARIABLE QUANTIFYING ITEM PAIRS IN REGARD TO AVERAGE DIFFICULTY AND DIFFICULTY DISCREPANCY	Little Difference_Moderate Difficulty (base)	1.0	
	Little Difference_High Difficulty	3.4**	[1.6,7.3]
	Little Difference_Low Difficulty	2.4***	[1.6,3.6]
	Moderate Difference_Lower Difficulty	1.5	[0.8,2.9]
	Moderate Difference_Higher Difficulty	3.4**	[1.5,7.6]
	High Difference	2.7	[1.0,7.5]
SIZE OF THE TESTLET	Both items from small testlets (base)	1.0	
	Both items from large testlets	0.4***	[0.3,0.7]
	One item from small with the second item from large testlet	0.8	[0.5,1.3]
ITEM FORMAT	Both items "Simple Multiple Choice" (base)	1.0	
	Both items ("Complex Multiple Choice" or "Short Response")	1.7	[0.8,3.4]
	Both items "Open Constructed Response"	1.6	[0.9,2.9]
	Pair of "Open Constructed Response" and "Simple Multiple Choice" items	0.4**	[0.2,0.7]
	Pair of "Simple Multiple Choice" and ("Complex Multiple Choice" or "Short Response") items	0.8	[0.5,1.3]
	Pair of "Open Constructed Response" and ("Complex Multiple Choice" or "Short Response") items	0.6	[0.3,1.2]

TYPE OF ITEM PAIRS	**Both Items Binary (base)**	**1.0**	
	One Item Binary One Polytomous	**0.4*** **	[0.2,0.6]
	Both Items Polytomous	**0.4**	[0.1,2.5]
READING TEXT TYPE	**Both items from "Non-Continuous Text (Map, Chart, Table or Form)" (base)**	**1.0**	
	Both items from "Continuous Text that is Narrative or Descriptive or Instruction"	**0.5**	[0.2,1.2]
	Both items from "Continuous Text that is Expository or Argumentative"	**0.5**	[0.2,1.0]
	Pair of "Continuous Text that is Narrative or Descriptive or Instruction" & Non-Continuous Text (Map, Chart, Table or Form)" items	**0.2*** **	[0.1,0.4]
	Pair of "Continuous Text that is Expository or Argumentative" & "Non-Continuous Text (Map, Chart, Table or Form)" items	**0.2*** **	[0.1,0.5]
	Pair of "Continuous Text that is Expository or Argumentative" & "Continuous Text that is Narrative or Descriptive or Instruction" items	**0.3** **	[0.1,0.7]
READING CONTEXT/SITUATION	**Both items Personal (base)**	**1.0**	
	Both items Public	**0.4***	[0.2,0.9]
	Both items Occupational	**1.0**	[0.4,2.4]
	Both items Educational	**0.3** **	[0.1,0.7]
	Pair of Personal and Public items	**0.6**	[0.3,1.1]
	Pair of Public and Occupational items	**0.4***	[0.2,0.9]
	Pair of Educational and Public items	**0.6**	[0.3,1.1]
	Pair of Occupational and Personal items	**0.7**	[0.4,1.5]
	Pair of Educational and Personal items	**0.8**	[0.4,1.5]
	Pair of Educational and Occupational items	**0.3***	[0.1,0.8]

** p<0.05, ** p<0.01 *** p<0.001 | Number of item pairs=11319*

The final model proposed consists of eight independent variables. The PISA study from 2009 produced increased odds 2.1 (1.4, 3.1) of finding positive LID among reading items compared to PISA 2000. In the year 2000 reading was the main domain tested, as it was again in 2009. This result suggests that a new batch of reading items introduced in 2009 was more positive LID prevalent. Bonferroni-corrected pairwise comparisons[39] also indicated marginally significant results when comparing PISA 2009 to PISA 2003 with odds equal to 4.9 (1, 23.9). This is unsurprising as reading the items used in PISA 2003 constituted a selection of the PISA 2000 item pool.

Results for the location of item pairs also proved to be as expected. The odds of finding positive LID for a pair of items from the same testlet as opposed to pairs from non-matching clusters were high and equal to 57.5 (24.1, 137,6). However, this effect size was not as large as the corresponding result for mathematics. This relationship between positive LID prevalence for mathematics and reading was indicated previously (Table 5.3.1). The influence of the relative difficulties of pairs of item resembles the corresponding results in mathematics with the exception that only marginal significantly (p=0.051) higher odds 2.7 (1.0, 7.5) of positive LID are observed for pairs of items which differ substantially (High difference) in their difficulty compared to the reference category. Controlling for all other variables in the final model, the odds of finding positive LID were reduced 0.4 (0.2, 0.7) when a pair of items came from large testlets[40] which was also the case in mathematics. Another result closely mimicking mathematics was found whether the item was coded binary or with partial credit scoring, compared to the base category of "Both items binary, " i.e. OR=0.4 (0.2, 0.6).

While considering item format, a pair of items where one question was "Open constructed response" and the other was "Simple Multiple Choice" reduced the odds of finding positive LID compared to the base of both items being "Simple Multiple Choice". Multiple testing Bonferroni, corrected for pairwise comparisons also suggested that the same setting of item pair (OCR and SMC) gives statistically significant odds ratios less than 1.0 compared to categories with both items being CMC/SR OR=0.2 (0.1, 0.7) or both items of OCR type OR=0.3 (0.1, 0.6).

The type of text showed reduced odds of finding positive LID in three item pairs involving non-matching text types compared to cases where both items were of non-continuous text type. Finally, situational classification of the item pairs indicated reduced odds of positive LID when both items were Educational and Public type as compared to the reference group of personally aimed pairs of questions.

[39] Not reported but available upon request
[40] Large meaning exceeding or equal to the median size of reading testlets which was 4 items

Science

Four additional variables specific to the science domain were included in the multilevel logistic regressions. Firstly, the scientific context variable represents combinations of items representing three situational contexts (*Personal, Global and Social*) (OECD, 2006, p. 27). Secondly, science competencies (*Using scientific evidence, Explaining phenomena scientifically, Identifying scientific issues*) (OECD, 2006, p. 29) are used in classifying pairs of items producing another variable. Thirdly, scientific application area was involved covering combinations of five topics (*Health, Natural resources, Hazards, Frontiers, and Environment*). The fourth variable used categories of scientific *Knowledge about science* (*Scientific enquiry and Scientific explanations)* and *Knowledge of science (Living systems, Physical and Technology systems, Earth and space systems)* as per the PISA 2006 framework (OECD, 2006, p. 32-33).

The Electronic Appendix for Table 5.4.3 - Science - Positive LID presents thirty-five models with the highlighted columns at the end of the file reporting the proposed final model along with its CIs. The final model is also given in Table 5.4.3. The AIC for final model was 1796 while BIC=1902. In comparison the null model had AIC=1872, BIC=1886. The proximity of BIC between null and final model suggests that the final model is not the most successful in explaining the presence of residual correlations larger than +0.1.

Table 5.4.3 Final multilevel logistic regression model predicting positive LID in science

		Model FINAL	95% CIs
PISA WAVE	PISA WAVE=2000 (base)	1.0	
	PISA WAVE=2003	1.1	[0.4,3.2]
	PISA WAVE=2006	1.4	[0.6,3.3]
	PISA WAVE=2009	1.8	[0.7,4.4]
	PISA WAVE=2012	1.2	[0.5,3.0]
LOCATION OF ITEM PAIRS	Items are not in the same cluster (base)	1.0	
	Items share the cluster but not testlet	0.1***	[0.0,0.3]
	Items are in the same testlet	5.0***	[2.4,10.1]
VARIABLE QUANTIFYING ITEM PAIRS IN REGARD TO AVERAGE DIFFICULTY AND DIFFICULTY DISCREPANCY	Little Difference_Moderate Difficulty (base)	1.0	
	Little Difference_High Difficulty	2.6**	[1.3,5.2]
	Little Difference_Low Difficulty	1.7*	[1.1,2.7]
	Moderate Difference_Lower Difficulty	3.8***	[1.9,7.4]
	Moderate Difference_Higher Difficulty	2.4	[1.0,5.7]
	High Difference	4.1	[0.6,26.2]
TYPE OF ITEM PAIRS	Both Items Binary (base)	1.0	
	One Item Binary One Polytomous	0.3**	[0.2,0.7]
	Both Items Polytomous	1.7	[0.1,21.8]

** p<0.05, ** p<0.01 *** p<0.001 | Number of item pairs=8600*

The final multilevel logistic model for positive LID in science is much smaller than the models for mathematics and reading, including only four variables. Similar to mathematic and contrary to the reading, no year effect was present for science.

The odds of finding positive LID were higher for pairs of questions from the same testlet compared to items from different clusters OR=5 (2.4, 10.1). This effect size is much smaller than is reported for other domains, suggesting that common testlet stimuli were not as influential for inducing positive LID as they are for the other cognitive domains.

Pairs of items that are similar to each other in terms of item difficulty produced slightly increased 2.6 (1.3, 5.2) and 1.7 (1.1, 2.7) and statistically significant odds of positive LID compared to the reference category of items with similar medium difficulties. Similarly, the odds of finding positive LID were 3.8 (1.9, 7.4) times higher for pairs of items that moderately differ from each other. Although these results are somewhat consistent with reading and mathematics, they are difficult to explain practically.

Item format is influential with odds of positive LID lower at 0.3 (0.2, 0.7) for pairs of mixed binary/polytomous items compared to the reference of both binarily coded questions.

5.4.2.2 Summary of quantitative investigation of positive LID drivers based on various PISA items characteristics

In summary, it was found that location of item pairs within the same testlet was the strongest predictor of positive LID in all cognitive domains, and more evident in mathematics and reading than in science. The quantitative results concurred with some of the regularities found in qualitative investigations based on the limited number of released items. The multilevel logistic regression for mathematics pointed to odds ratios exceeding 4 when geometrically or statistically related items were looked at in comparison to the reference category - “Number”. This also agreed with some qualitative observations. An increased likelihood of positive dependency when items differed in their difficulties is noted, and it was speculatively suggested that external assistance might be one possible explanation for this specific to mathematics result, although other possible explanations should be explored. Items belonging to larger testlets were found to reduce the odds of positive LID for mathematics and for reading. A final model for science was smaller in comparison to mathematic or reading investigations. As suggested in the qualitative science domain investigations, the possibility of positive dependency among items from science competency of “Identifying scientific issues” did not emerge in the final model, but it was present in early models as reported in the corresponding electronic appendix.

5.4.2.1 Model explaining negative LID

Mathematics

Multilevel logistic regressions predicting negative LID, i.e. residual correlation less than -0.1, used the same set of variables as reported in the previous section showing results for mathematics and positive dependency. The Electronic Appendix for Table 5.4.4 - Mathematics - Negative LID presents forty-three models. The final model is given in Table 5.4.4. Null and final models produced AICs equal to 4864 and 4465, respectively, along with BICs equal to 4878 and 4910, respectively. As the final model includes all but three variables available, the final model is not the most parsimonious that is reflected in BIC information criteria which is conservative in applying a penalty term for the number of parameters in the model.

Table 5.4.4 Final multilevel logistic regression model predicting negative LID in mathematics

		Model FINAL	95% CIs
PISA WAVE	PISA WAVE=2000 (base)	1.0	
	PISA WAVE=2003	0.7	[0.4,1.2]
	PISA WAVE=2006	0.6	[0.4,1.1]
	PISA WAVE=2009	1.3	[0.7,2.3]
	PISA WAVE=2012	0.5*	[0.3,0.9]
LOCATION OF ITEM PAIRS	Items are not in the same cluster (base)	1.0	
	Items share the cluster but not testlet	0.1***	[0.1,0.2]
	Items are in the same testlet	(not estimable)[41]	
VARIABLE QUANTIFYING ITEM PAIRS IN REGARD TO AVERAGE DIFFICULTY AND DIFFICULTY DISCREPANCY	Little Difference_Moderate Difficulty (base)	1.0	
	Little Difference_High Difficulty	1.0	[0.7,1.4]
	Little Difference_Low Difficulty	1.2	[0.8,1.6]
	Moderate Difference_Lower Difficulty	2.5***	[1.7,3.5]
	Moderate Difference_Higher Difficulty	3.1***	[2.2,4.3]
	High Difference	7.4***	[4.9,11.1]
SIZE OF THE TESTLET	Both items from small testlets (base)	1.0	
	Both items from large testlets	1.8***	[1.3,2.5]
	One item from small with the second item from large testlet	1.5***	[1.2,1.9]
ITEM FORMAT	Both items "Simple Multiple Choice" (base)	1.0	
	Both items ("Complex Multiple Choice" or "Short Response")	1.7	[1.0,3.0]
	Both items "Open Constructed Response"	2.1*	[1.1,4.2]
	Pair of "Open Constructed Response" and "Simple Multiple Choice" items	1.3	[0.7,2.4]
	Pair of "Simple Multiple Choice" and ("Complex Multiple Choice" or "Short Response") items	1.4	[0.8,2.4]
	Pair of "Open Constructed Response" and ("Complex Multiple Choice" or "Short Response")	1.9*	[1.1,3.4]
TYPE OF ITEM PAIRS	Both Items Binary (base)	1.0	
	One Item Binary One Polytomous	0.6***	[0.4,0.7]
	Both Items Polytomous	0.3	[0.1,1.0]
INSTITUTIONAL SOURCE OF ITEM PAIRS	Institutional source not agree (base)	1.0	
	Institutional source agree	1.6***	[1.3,2.1]

[41] It was found that for reading and science domains there was not a single item pair from the within-testlet for which residual correlations were lower than -0.1. For the sake of consistency in mathematics a single item pair from testlet M998 was removed. Default approach of melogit procedure was retained which excluded all within-testlet item pairs.

ITEMS LENGTH **Short questions contain fewer than 50 words. Medium-length questions contain 51 to 100 words. Long questions contain more than 100 words (OECD, 2010b)**	**Both items Medium length (base)**	**1.0**	
	Both items Long length	**0.3*** **	[0.2,0.5]
	Both items Short length	**0.4*** **	[0.2,0.6]
	Pair of Long and Medium length items	**0.5*** **	[0.4,0.7]
	Pair of Short and Medium length items	**0.5*** **	[0.3,0.7]
	Pair of Short and Long length items	**0.4*** **	[0.2,0.5]
	Length for at least one item is missing	**0.2** **	[0.1,0.5]
MATHEMATICAL PROCESS	**Both items Interpret (base)**	**1.0**	
	Both items Employ	**1.5**	[0.9,2.6]
	Both items Formulate	**2.2** **	[1.3,4.0]
	Pair of Formulate and Employ items	**1.9** **	[1.2,3.2]
	Pair of Formulate and Interpret items	**1.6**	[0.9,2.6]
	Pair of Interpret and Employ length items	**1.7***	[1.1,2.8]
	Competency for at least one item is missing	**2.6** **	[1.4,4.8]
"BRANCH" OF MATHEMATICS	**Both items Number (base)**	**1.0**	
	Both items Algebra/Functions/Discrete Mathematics	**2.2**	[0.9,5.5]
	Both items Geometry	**0.7**	[0.2,2.1]
	Strand for at least one item is missing	**4.2***	[1.3,13.2]
	Both items Statistics/Probability/Data	**1.0**	[0.5,1.8]
	Both items Measurement	**0.7**	[0.1,3.7]
	Pair of Algebra/Functions/Discrete Mathematics and Geometry items	**2.0***	[1.1,3.8]
	Pair of Algebra/Functions/Discrete Mathematics and Number items	**0.7**	[0.4,1.3]
	Pair of Algebra/Functions/Discrete Mathematics and Statistics/Probability/Data items	**2.1** **	[1.2,3.6]
	Pair of Algebra/Functions/Discrete Mathematics and Measurement items	**1.3**	[0.6,2.9]
	Pair of Geometry and Number items	**1.6**	[1.0,2.6]
	Pair of Geometry and Statistics/Probability/Data items	**1.7***	[1.0,2.9]
	Pair of Geometry and Measurement items	**0.8**	[0.4,1.8]
	Pair of Number and Statistics/Probability/Data items	**1.3**	[0.8,2.0]
	Pair of Number and Measurement items	**1.3**	[0.8,2.3]
	Pair of Statistics/Probability/Data and Measurement items	**1.6**	[0.9,2.8]
MATHEMATICAL CONTEXT/SITUATION	**Both items Personal (base)**	**1.0**	
	Both items Public	**1.3**	[0.7,2.3]
	Both items Scientific	**0.5**	[0.3,1.1]
	Both items Occupational	**0.1** **	[0.0,0.6]
	Both items Educational	**2.1**	[0.7,5.8]

Pair of Scientific and Public items	**1.2**	[0.7,2.2]
Pair of Scientific and Personal items	**0.8**	[0.4,1.5]
Pair of Scientific and Occupational items	**0.9**	[0.5,1.8]
Pair of Educational and Scientific items	**1.3**	[0.6,2.7]
Pair of Personal and Public items	**1.2**	[0.7,2.2]
Pair of Public and Occupational items	**0.9**	[0.5,1.7]
Pair of Educational and Public items	**1.3**	[0.6,2.6]
Pair of Occupational and Personal items	**0.7**	[0.3,1.4]
Pair of Educational and Personal items	**0.6**	[0.3,1.3]
Pair of Educational and Occupational items	**0.2***	[0.0,0.7]

** p<0.05, ** p<0.01 *** p<0.001 | Number of item pairs=8663*

While the variable representing year produced no significant result for any PISA instances as compared to the initial PISA 2000 wave, Bonferroni corrected pairwise comparisons showed increased odds of finding RCs<-0.1 for PISA 2009 compared to PISA 2003 with OR=1.8 (1.1, 3) and PISA 2006 2.1 (1.2, 3.8).

The variable quantifying the difficulty and relative location of item pairs returned an odds ratio of 7.4 (4.9, 11.1) for items that were on the opposite range of difficulty difference (exceeding the difference of 60% correctness as visualised in Figure 5.4.17), compared to the base reference category. Similarly, statistically significant odds ratios >1.0 are visible for both categories involving items that are of moderate difference in difficulty as expressed by OR=2.5 (1.7, 3.5) and OR=3.1 (2.2, 4.3). If selective time and effort allocation against the items appearing difficult was, as suggested by Yen (1993), one of the negative dependency drivers, the results listed above would be expected. Furthermore, Bonferroni corrected pairwise comparisons produced significant odds >2 of negative LID for the three categories featured above, in comparison to both "Little difference" levels of this variable.

The same explanation of selective effort allocation could be speculated while interpreting the results for variable labelled "Item Formats". Controlling for other variables in the model, there were higher odds of 2.1 (1.1, 4.2) and 1.9 (1.1, 1.9) of finding negative dependency among item pairs which had at least one item of "Open constructed response" type as compared to a base reference of both items being "Simple multiple choice". A student could selectively not put effort into questions, which in their judgement, would require extra effort in writing an answer and justifying it. The same plausible explanation could be behind the results for the variable discussing mathematical processes. Pairs of items for which at least one questions was more cognitively demanding, using the mathematical process of "Formulate", show significantly higher than 1 odds of negative LID, compared to the reference category of both questions being of "Interpret" type. Also, a variable investigating the mathematical strand produced higher than 1 odds for some categories featuring "Geometry" and "Statistics/Probability/Data" items. Finally, the odds of finding residual correlations less than -0.1 were higher - 1.8 (1.3, 2.5) for item pairs from non-matching larger testlets, compared to item pairs from small testlets.

A variable which quantified the origins of the items produced odds of negative LID equal to 1.6 (1.3, 2.1) for items pairs from the same source as compared to reference category of item pair from different sources for submission for PISA use. No plausible explanation for this result can be offered, but this variable could be acting as a proxy for other non-measured factors.

Four variables reported odds ratios of negative dependency being significantly smaller than one

while controlling for other variables in the model. Firstly, a predictor quantifying the length of the item pairs reported such odds in reference to pairs of questions that both were of a medium length. Secondly, the same trend was present in a variable looking at the location of the items within the same cluster versus reference of items not being located within the same cluster. Thirdly, situational context of the item pairs presented reduced odds for some categories involving the "Occupational" type of questions, compared to both items being Personal. Fourthly, as was the case for models predicting positive LID, the item pairs from different item format types showed an odds ratio of 0.6 (0.4, 0.7) in relation to the reference of both items being of binary type. The limited literature regarding negative dependency does not offer plausible explanations for these results, neither do the results of qualitative investigations in section 5.4.1.

Reading

In the quantitative investigation of predictors of negative LID, the same set of eleven variables utilised in positive LID modelling was used. The Electronic Appendix for Table 5.4.5 - Reading - Negative LID presents thirty-one models organised in the fashion described at the beginning of Section 5.4.2. The final model is given in Table 5.4.5. The final model's AIC was 6418 and BIC was 6705, while the respective statistics for the null model were 7180 and 7195.

Table 5.4.5 Final multilevel logistic regression model predicting negative LID in reading

		Model FINAL	95% CIs
PISA WAVE	PISA WAVE=2000 (base)	1.0	
	PISA WAVE=2003	4.1***	[2.6,6.4]
	PISA WAVE=2006	1.8*	[1.0,3.2]
	PISA WAVE=2009	3.5***	[2.9,4.2]
	PISA WAVE=2012	7.0***	[5.1,9.5]
LOCATION OF ITEM PAIRS	Items are not in the same cluster (base)	1.0	
	Items share the cluster but not testlet	0.1***	[0.1,0.1]
	Items are in the same testlet	(not estimable)[42]	
VARIABLE QUANTIFYING ITEM PAIRS IN REGARD TO AVERAGE DIFFICULTY AND DIFFICULTY DISCREPANCY	Little Difference_Moderate Difficulty (base)	1.0	
	Little Difference_High Difficulty	0.9	[0.5,1.4]
	Little Difference_Low Difficulty	2.2***	[1.8,2.6]
	Moderate Difference_Lower Difficulty	2.9***	[2.2,3.7]
	Moderate Difference_Higher Difficulty	1.4	[0.9,2.1]
	High Difference	3.3***	[2.1,5.2]
SIZE OF THE TESTLET	Both items from small testlets (base)	1.0	
	Both items from large testlets	1.7***	[1.3,2.2]
	One item from small with the second item from large testlet	1.2	[1.0,1.5]
ITEM FORMAT	Both items "Simple Multiple Choice" (base)	1.0	
	Both items ("Complex Multiple Choice" or "Short Response")	1.1	[0.7,1.6]
	Both items "Open Constructed Response"	2.1***	[1.5,2.8]
	Pair of "Open Constructed Response" and "Simple Multiple Choice" items	1.9***	[1.4,2.4]
	Pair of "Simple Multiple Choice" and ("Complex Multiple Choice" or "Short Response") items	1.3*	[1.0,1.7]
	Pair of "Open Constructed Response" and ("Complex Multiple Choice" or "Short Response") items	2.1***	[1.5,2.7]

[42] It was found that for reading and science domains there was not a single item pair from the within-testlet for which residual correlations were lower than -0.1. Default approach of melogit procedure was retained which excluded all within-testlet item pairs.

TYPE OF ITEM PAIRS	**Both Items Binary (base)**	1.0	
	One Item Binary One Polytomous	0.7**	[0.5,0.9]
	Both Items Polytomous	0.6	[0.2,1.8]
LANGUAGE FAMILIES IN WHICH ITEM PAIRS WERE SUBMITTED	**Both items submitted in English (base)**	1.0	
	English item and Hellenic or Italic item	1.2	[0.9,1.5]
	English item and Germanic family item	1.0	[0.8,1.3]
	English item and Japanese or Korean item	0.7	[0.5,1.0]
	English item and Other (Czech, Finnish or Hungarian)	1.1	[0.9,1.5]
	Both items from Hellenic or Italic family	1.9***	[1.3,2.8]
	Both items from Germanic family	1.4	[0.8,2.3]
	Both items from Japanese or Korean language	1.2	[0.5,2.9]
	Both items Other (Czech, Finnish or Hungarian)	1.1	[0.5,2.6]
	Other combination of two non-English items	1.1	[0.8,1.4]
READING CONTEXT/SITUATION	**Both items Personal (base)**	1.0	
	Both items Public	1.3	[0.9,1.9]
	Both items Occupational	0.8	[0.4,1.5]
	Both items Educational	1.2	[0.8,1.8]
	Pair of Personal and Public items	1.3	[0.9,1.8]
	Pair of Public and Occupational items	1.7**	[1.1,2.4]
	Pair of Educational and Public items	1.4	[1.0,1.9]
	Pair of Occupational and Personal items	1.5*	[1.1,2.2]
	Pair of Educational and Personal items	1.2	[0.9,1.6]
	Pair of Educational and Occupational items	1.2	[0.9,1.8]

** $p<0.05$, ** $p<0.01$ *** $p<0.001$ | Number of item pairs=11650*

Odds ratios for finding negative dependency in reading for PISA 2003, 2006, 2009 and 2012 were above 1 with the PISA 2000 wave as a reference. Bonferroni corrected pairwise comparisons additionally showed an odds ratio of 3.8 (1.6, 9.1) and 2.0 (1.4, 2.8) for PISA 2012, compared to PISA 2006 and 2009, respectively. PISA 2000 and PISA 2009 had reading literacy as the main targeted domain, and in qualitative investigations, it was observed that targeted domains produce less negative dependency.

As in mathematics pairs of items, at opposite ends of difficulty levels, reported increased odds of negative LID of 3.3 (2.1, 5.2) in reference to item pairs of moderate difficulty and similar complexity. Also, similar to mathematical results, three categories of the "Item format" variable reported odds of negative dependency close to 2 for pairs of reading items involving at least one "Open constructed response" question with the reference category being "Simple multiple choice" item pairs. It is argued that these results may indicate selective effort allocation as stipulated by Yen (1993).

Three other variables produced statistically significant results with odds exceeding 1 for which justifications are hard to provide. Odds of negative LID were 1.9 for items which were submitted using a Hellenic or Italic family of language, compared to both items being originally designed and submitted in English. Perhaps some challenges in translations of differently sourced items (Arffman, 2010) could be partly responsible. The variable quantifying item pairs' situational context produced odds ratios >1.0 for reading item pairs involving "Occupational" items, compared to a reference of "Personal" items. Finally, a variable describing the size of the testlet from which items originated, reported the odd ratio of 1.7 (1.3, 2.2) when both items came from large testlets, compared to both items being from small testlets. Access to all PISA cognitive items could allow investigating this result particularly in relation to the variable representing language in which the reading items were submitted to PISA.

Science

The final negative dependency model for science is presented in Table 5.4.6, while the remaining analyses leading to it are reported in the Electronic Appendix for Table 5.4.6 - Science - Negative LID. AIC and BIC for the model reported in the Table 5.4.6 were 3258 and 3646, respectively. Corresponding null model values were AIC=3540, BIC=3555. The BIC larger for the final model as compared to null model is a consequence of using variables with many categories.

Table 5.4.6 Final multilevel logistic regression model predicting negative LID in science

		Model FINAL	95%CIs
PISA WAVE	**PISA WAVE=2000 (base)**	1.0	
	PISA WAVE=2003	1.8*	[1.0,3.3]
	PISA WAVE=2006	0.9	[0.5,1.4]
	PISA WAVE=2009	1.6	[0.9,2.7]
	PISA WAVE=2012	1.6	[1.0,2.7]
LOCATION OF ITEM PAIRS	**Items are not in the same cluster (base)**	1.0	
	Items share the cluster but not testlet	0.0***	[0.0,0.0]
	Items are in the same testlet	(not estimable)[43]	
VARIABLE QUANTIFYING ITEM PAIRS IN REGARD TO AVERAGE DIFFICULTY AND DIFFICULTY DISCREPANCY	**Little Difference_Moderate Difficulty (base)**	1.0	
	Little Difference_High Difficulty	1.1	[0.7,1.7]
	Little Difference_Low Difficulty	1.4*	[1.0,1.9]
	Moderate Difference_Lower Difficulty	1.7*	[1.1,2.6]
	Moderate Difference_Higher Difficulty	1.3	[0.7,2.3]
	High Difference	3.1*	[1.3,7.9]
SIZE OF THE TESTLET	**Both items from small testlets (base)**	1.0	
	Both items from large testlets	1.1	[0.7,1.9]
	One item from small with the second item from large testlet	1.3	[1.0,1.6]
ITEM FORMAT	**Both items "Simple Multiple Choice" (base)**	1.0	
	Both items ("Complex Multiple Choice" or "Short Response")	1.6	[0.9,2.6]
	Both items "Open Constructed Response"	1.5	[0.9,2.4]
	Pair of "Open Constructed Response" and "Simple Multiple Choice" items	1.9**	[1.3,2.8]
	Pair of "Simple Multiple Choice" and ("Complex Multiple Choice" or "Short Response") items	1.9**	[1.3,2.8]
	Pair of "Open Constructed Response" and ("Complex Multiple Choice" or "Short Response") items	1.5*	[1.0,2.4]
LANGUAGE FAMILIES IN WHICH ITEM PAIRS WERE SUBMITTED	**Both items submitted in English (base)**	1.0	
	English item and Hellenic or Italic item	1.4	[0.8,2.5]
	English item and Germanic family item	2.3**	[1.4,3.7]
	English item and Japanese or Korean item	2.8***	[1.5,5.2]
	Both items from Hellenic or Italic family	0.9	[0.3,2.2]
	Both items from Germanic family	2.0**	[1.2,3.3]
	Both items from Japanese or Korean language	6.9***	[2.3,20.6]
	Other combination of two non-English items	1.2	[0.7,2.1]

[43] It was found that for reading and science domains there was not a single item pair from the within-testlet for which residual correlations were lower than -0.1. Default approach of melogit procedure was retained which excluded all within-testlet item pairs.

INSTITUTIONAL SOURCE OF ITEM PAIRS	**Institutional source not agree (base)**	1.0	
	Institutional source agree	1.8**	[1.2,2.6]
SCIENCE APPLICATION AREA	**Both items Health (base)**	1.0	
	Both items Natural resources	0.7	[0.3,1.5]
	Both items Hazards	0.4	[0.2,1.2]
	Both items Frontiers	0.5*	[0.2,1.0]
	Both items Environment	0.6	[0.3,1.1]
	Pair of Health and Hazards items	0.7	[0.4,1.3]
	Pair of Health and Frontiers items	0.6*	[0.3,1.0]
	Pair of Natural resources and Health items	0.5*	[0.3,0.9]
	Pair of Natural resources and Hazards items	0.6	[0.3,1.2]
	Pair of Natural resources and Frontiers items	0.2***	[0.1,0.5]
	Pair of Natural resources and Environment items	0.5*	[0.3,0.9]
	Pair of Health and Environment items	0.8	[0.5,1.4]
	Pair of Hazards and Frontiers items	0.3**	[0.2,0.7]
	Pair of Hazards and Environment items	0.5*	[0.3,0.9]
	Pair of Frontiers and Environment items	0.5*	[0.3,0.9]
SCIENCE KNOWLEDGE	**Both items Knowledge of science - Living systems (KoS_LS) (base)**	1.0	
	Both items Knowledge of science - Physical and Technology systems (KoS_PaTS)	5.4***	[2.1,13.9]
	Both items Knowledge of science - Earth and space systems (KoS_EaSS)	4.5*	[1.4,14.5]
	Both items Knowledge about science - Scientific explanations (KaS_SEXP)	3.6**	[1.5,8.5]
	Both items Knowledge about science - Scientific enquiry (KaS_SENQ)	1.4	[0.5,3.7]
	Pair of KoS_PaTS and KoS_LS items	2.2	[0.9,5.3]
	Pair of KoS_PaTS and KoS_EaSS items	3.8**	[1.5,9.4]
	Pair of KoS_PaTS and KaS_SEXP items	3.4**	[1.5,7.9]
	Pair of KoS_PaTS and KaS_SENQ items	3.0*	[1.3,7.1]
	Pair of KoS_LS and KoS_EaSS items	2.1	[0.8,5.2]
	Pair of KoS_LS and KaS_SEXP items	2.6*	[1.1,5.9]
	Pair of KoS_LS and KaS_KaS_SENQ items	1.4	[0.6,3.4]
	Pair of KoS_EaSS and KaS_SEXP items	3.4**	[1.4,8.1]
	Pair of KoS_EaSS and KaS_KaS_SENQ items	2.0	[0.8,5.1]
	Pair of KaS_SEXP and KaS_KaS_SENQ items	2.6*	[1.1,6.0]

** p<0.05, ** p<0.01 *** p<0.001 | Number of item pairs=8589*

The science investigations regarding negative LID produced very similar results to the corresponding mathematical and reading models in regard to predictors involving formats of the items and the relative difficulty of item pairs. It is argued that selective effort allocation may also be at play in investigating item pairs from different combinations of science knowledge categories. All categories of this variable involving "Knowledge about science - Scientific explanations (KaS_SEXP)" gave odds ratios >1.0 with the reference being pairs of items requiring knowledge of living systems. The largest odds ratio 5.4 (2.1, 13.9) came, however when a pair of items from "Knowledge of science - Physical and Technology systems (KoS_PaTS)" is compared to the reference category. Perhaps students allocated less time and effort to some types of scientific questions which may be driving the negative LID results.

Another result worth highlighiting relates to the variable quantifying the language origin of items with negative dependency being more likely for scientific items submitted in Japanese or Korean showing odds ratio of 6.9 (2.3, 20.6) in comparison to the reference category of both items being of English origin. Given the effort which is given by PISA to the quality of language translations used in PISA, it is likely that specific curriculum preferences are involved in this results. The variable quantifying a "Science application area" produced odds ratios <1.0 for categories involving "Frontiers", compared to the base reference of both items being of "Health" science application.

5.4.2.2 Summary of quantitative investigation of negative LID drivers based on various PISA items characteristics

Habing and Roussos (2003) suggested that negative dependency is a mathematical consequence of the existence of positive LID. While qualitative investigations and the graphical visualisations associated with them pointed towards this conclusion, the cross domain consistency in some multilevel logistic regressions strengthens the argument that part of the negative LID could be due to selective time and effort allocation as suggested by Yen (1993). This point is supported by the consistency of the results involving item pairs with high difficulty difference or involving the need to write answers to "Open constructed response" types of cognitive questions. Some of the results of models predicting negative LID were difficult to explain, and it emerged from qualitative investigations that the models predicting negative dependency might benefit from the inclusion of additional variables quantifying the placement of item pairs in a two hour long testing regime. This limitation is additionally supported by recent publications by Bolsinova, Tijmstra, Molenaar, and De Boeck (2017) and Bolsinova, de Boeck, and Tijmstra (2017) which suggest that the response time can contribute to negative dependency.

CHAPTER 6 RESULTS FOR RESEARCH AIM 3 - LID IN THE PISA'S NATIONAL CALIBRATIONS

6.1 The organisation of the chapter

This chapter is divided into two main sections. Section 6.2 systematically looks at the prevalence of LID in national cognitive datasets and is further organised into six parts looking at positive and negative LID for mathematics, reading and science. Each part offers consistency in the flow of the LID investigation where graphs reporting LID prevalence are followed by analyses identifying countries with unusually high levels of dependency. The primary aim of Section 6.3 is to consider the reporting of cross-country and cross-wave consistency in LID. Once again six sub-sections are presented to report on pairs of within-testlet and between-testlet items for the three cognitive domains. Readers interested in the specific type of LID and cognitive domains are encouraged to use pdf bookmarks to facilitate targeted navigation of the chapter.

6.2 LID prevalence at national level calibrations level data pointing to economies with high levels of dependency

This section aims to report the prevalence of residual correlations exceeding a cut-point of 0.1 as a percentage, in a similar manner to Table 5.2.2, reported for international calibrations in section 5.2. It also highlights the countries which may have a higher LID prevalence. Research questions RQ_3A and RQ_3B are addressed in this subchapter.

6.2.1 National calibrations with positive LID in mathematics

Due to the involvement of 24 national calibrations across three cognitive domains and five PISA waves, the results cannot be reported in easy to view and effective text form that mimics the reporting format used in Table 5.2.2. Consequently, Figure 6.2.1 offers a graphical overview of the percentage of RCs exceeding a cut-point of 0.1 (blue markers) in each of 24 OECD countries from the mathematical domain perspective. This figure is presented as a dual graph showing simultaneously the number of students involved in CFAs estimations (green markers). Figure 6.2.2 extends Figure 6.2.1 by introducing the distribution of positive LID prevalence to indicate components due to within-testlet (dark blue) and between-testlet (light blue) locations of items' pairs.

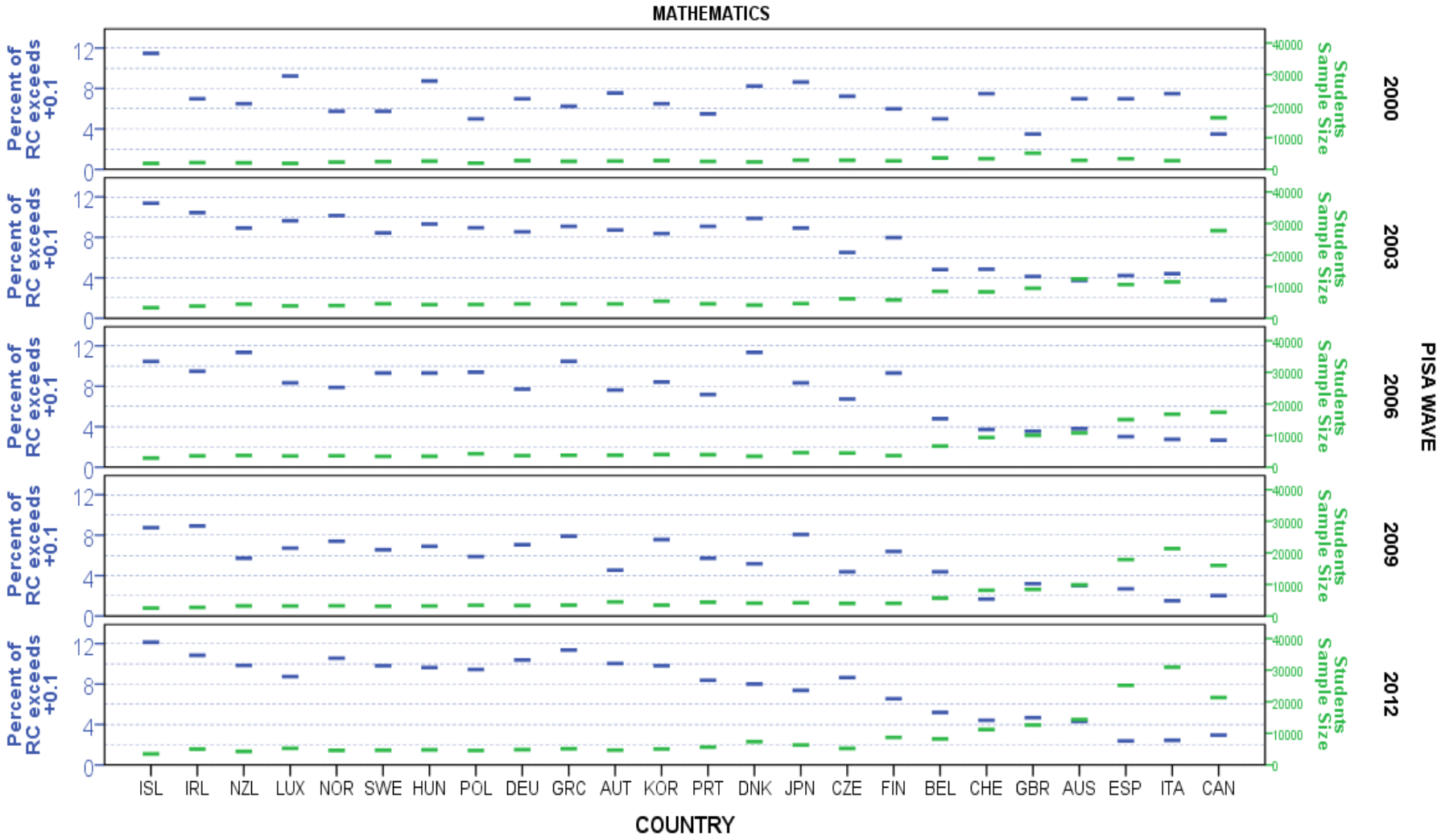


Figure 6.2.1 Dual graph showing the percent of mathematics item pairs with RCs exceeding 0.1 taken out of all RCs against students' sample sizes

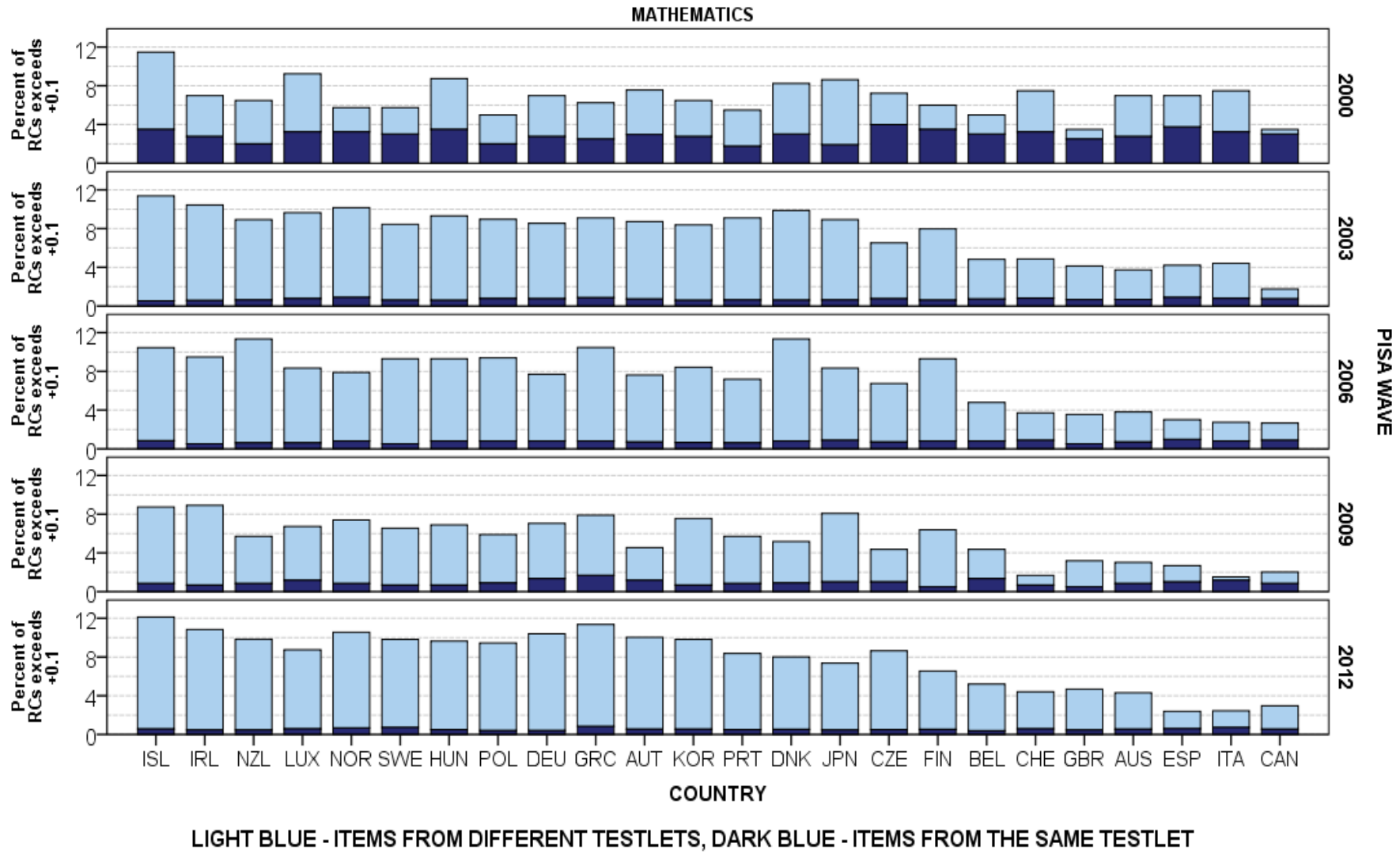


Figure 6.2.2 Percent of mathematics item pairs with RCs exceeding 0.1 taken out of all RCs separated into components involving item pairs from the same testlets and from different testlets

The concerns expressed in the limitation section 7.2.1 regarding the tenability of using a fixed 0.1 cut-point can be clearly seen in Figure 6.2.1 where the percent of RCs exceeding 0.1 negatively covary with the number of students employed in each PISA national implementation. Figure 6.2.2 offers an extension to this statement, pointing out that the dependence on a students' sample size seems to be related primarily to the positive RCs involving pairs of items from different testlets. At the same time, there are differences in the prevalence for countries that used similarly sized student's samples. For example, in PISA 2006, Portugal (PRT) reported the lower percentage[44] of RCs above 0.1 cut point (7.2%) as compared to (11.3%) from Denmark (DNK), despite countries using comparable students' sample sizes. Similarly, in PISA, 2012 Greece's (GRC) prevalence of 11.4% was higher[45] as compared to a similarly sized cohort of students from Luxembourg (LUX) 8.7%. The comparisons mentioned above are limited only to comparing countries that approached a similar number of students. The proportional allocation of cut point 0.1 exceeding RCs due to item pairs' allocation also vary. It can be inferred from Figure 6.2.2 that in PISA 2000 Finland's (FIN) within-testlet high RCs were 1.4 times more predominant as compared to the between-testlets high RCs (3.5% versus 2.5%). On the contrary Japanese (JAP) results were reversed with positive LID indicating pairs of items from different testlets being approximately 3.5 times more prevalent as compared to within-testlet results (6.7% versus 1.9%). The students' sample sizes in both countries were quite comparable differing by only 200 students. These descriptive results contributed to an investigation of outlying countries in regard to two types of positive LID (between-testlets and within-testlet). The cross wave comparison of results presented in Figure 6.2.2 is questionable due to a different number of items being used from study to study and also a difference in the arrangements of testlets' sizes. However, prominently larger proportions of within-testlet positive LID in PISA 2000 are likely to be partly because of a high proportion of four items large testlets used in this year as compared to subsequent waves (see Figure 4.2.1 for reference).

In order to address the issue of the negative relation of prevalence estimates to the student sample sizes, the countries with higher levels of positive LID will be identified after fitting the reciprocal function. The choice of function was driven by the figures from a publication

[44] Non-corrected for multiple comparisons 95%CIs for the proportion difference is (1.7%-6.5%). Epitools were used - http://epitools.ausvet.com.au/content.php?page=z-test-2&p1=0.072&p2=0.113&n1=1128&n2=1128&Conf=0.05&tails=2&samples=2

[45] Non-corrected for multiple comparisons 95%CIs for the proportion difference is (1.3%-4.1%). Epitools were used - http://epitools.ausvet.com.au/content.php?page=z-test-2&p1=0.087&p2=0.114&n1=3486&n2=3486&Conf=0.05&tails=2&samples=2

by Christensen et al. (2017) which has been elaborated in the methodology chapter (Section 3.4.1). Table 6.2.1 assisted in highlighting countries with outlying high positive LID involving pairs of items from different testlets, and it reports various tests for detecting prediction residuals outliers. In that table, three alternative methods for identifying cases as outliers – namely Grubb's Single Outlier Test, Rosner's procedure, and Tukey's Outside Values Test – are used. The prediction limits along with Box-and-Whisker plots of prediction residuals are shown in Figures 6.2.3-6.2.7.

Table 6.2.1 Countries with high levels of between-testlets positive LID in mathematics

	R2 FOR Y=1/(A+BX)	COUNTRY	Value of Possible Outlier	ESD |Z|	Grubbs' Single-Outlier Level Test Prob Level (Alternative Hypothesis: One-Sided vs Maximum)	Conclude Outlier by Rosner's Procedure	Tukey, 1977 - Outside values
PISA 2000	0.42	**Japan (JPN)**	3.1	2.46	0.104	No	**Yes**
		Iceland (ISL)	2.5	2.41	0.117	No	No
PISA 2003	0.98	**Korea (KOR)**	0.9	2.31	**0.174**	**Yes**	No
		Finland (FIN)	0.9	2.66	**0.043**	**Yes**	No
PISA 2006	0.87	**New Zealand (NZL)**	2.5	2.44	0.110	No	**Yes**
		Denmark (DNK)	1.8	2.14	0.287	No	No
PISA 2009	0.84	**Japan (JPN)**	2.4	2.69	**0.041**	**Yes**	**Yes**
		Korea (KOR)	1.2	1.71	0.915	No	No
PISA 2012	0.95	**Greece (GRC)**	1.7	2.50	**0.087**	**Yes**	**Yes**
		Ireland (IRL)	1.4	2.50	**0.081**	**Yes**	No

Grey areas are used to highlight PISA waves for which mathematics was a targeted domain.

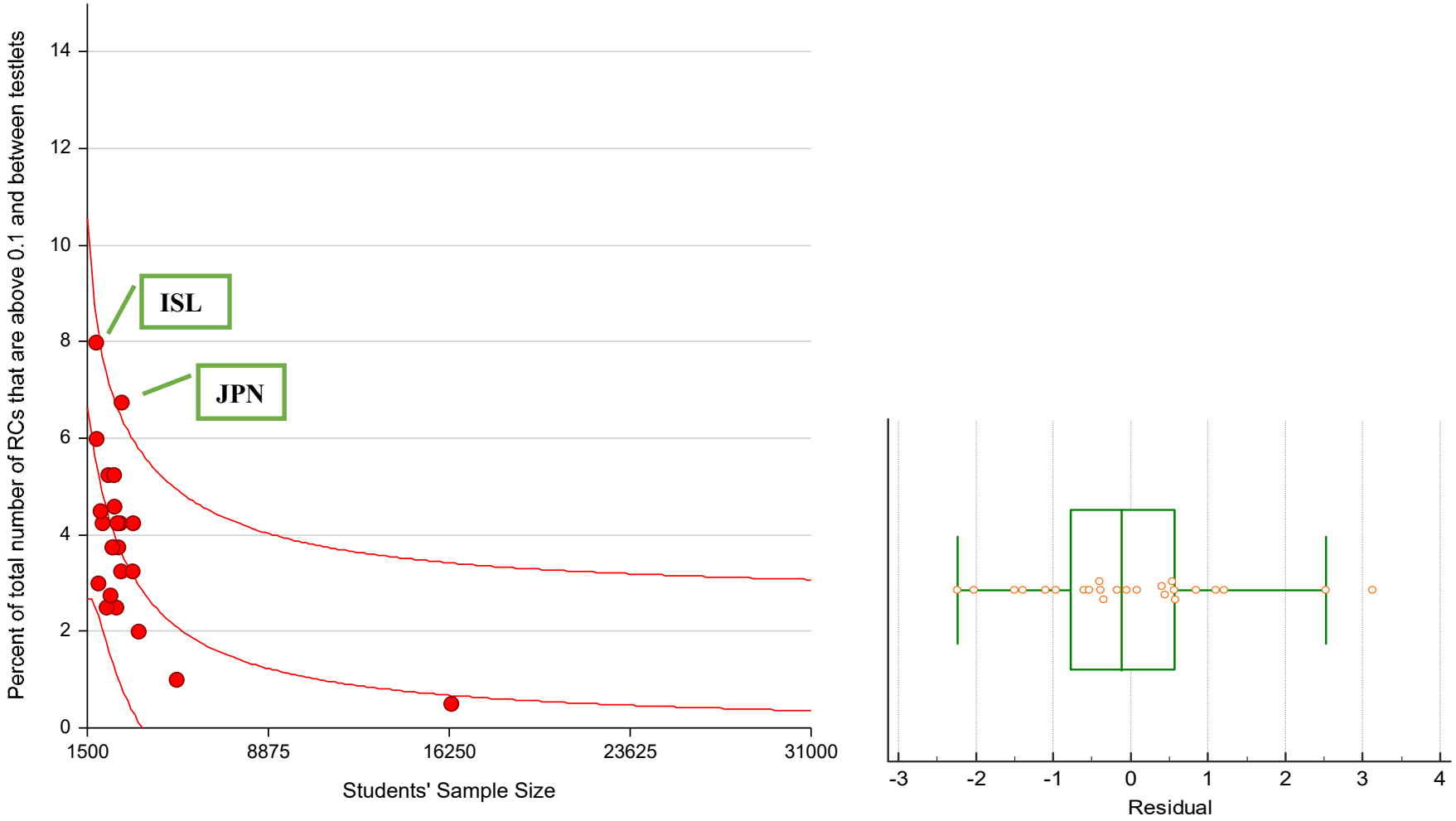


Figure 6.2.3 Reciprocal function and its prediction limits fitted to picture association between student sample size and prevalence of RCs along with a box-plot of prediction residuals (PISA 2000 / RCs that are above 0.1 / Pairs of items from different testlets / Mathematics)

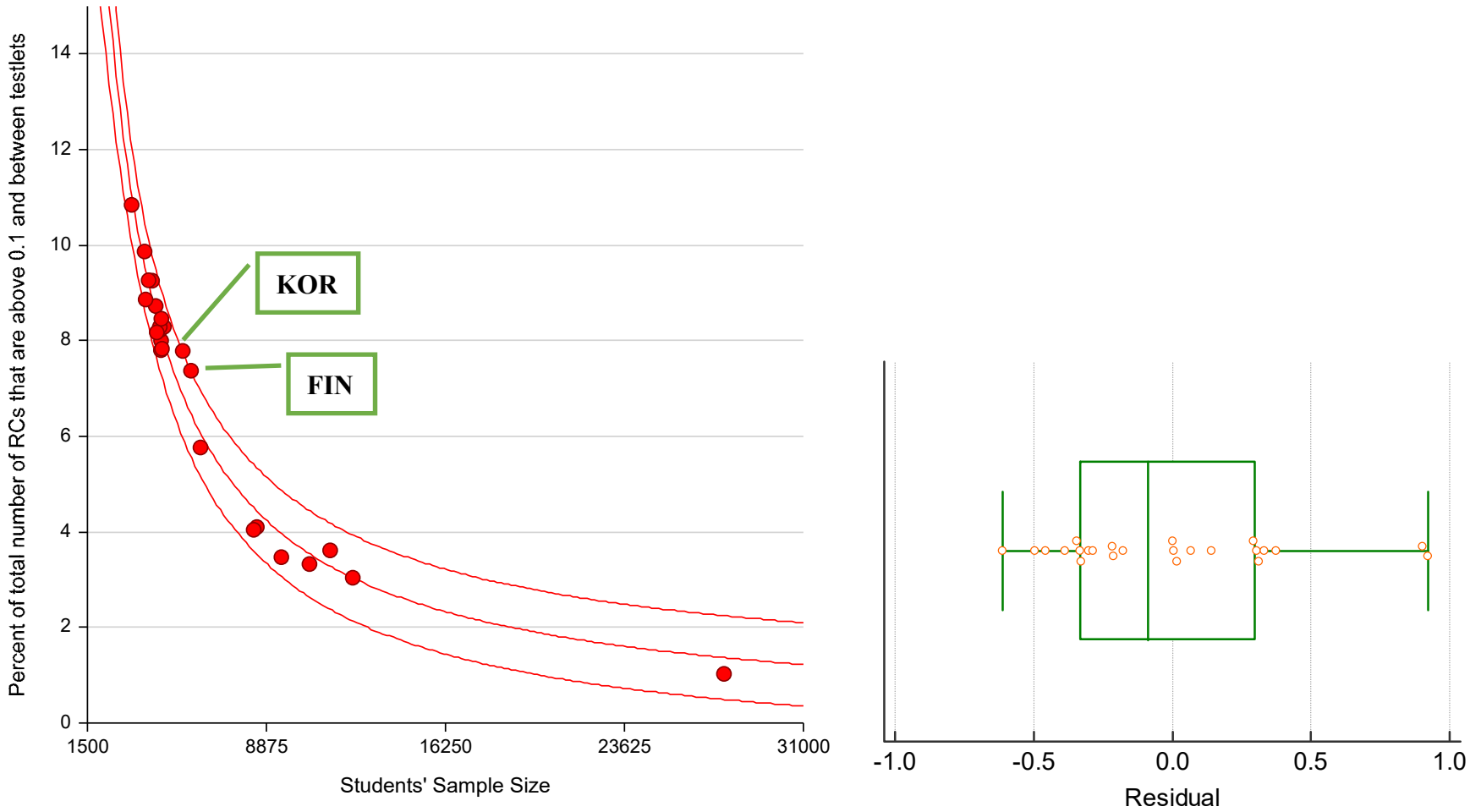


Figure 6.2.4 Reciprocal function and its prediction limits fitted to picture association between student sample size and prevalence of RCs along with a box-plot of prediction residuals (PISA 2003 / RCs that are above 0.1 / Pairs of items from different testlets / Mathematics)

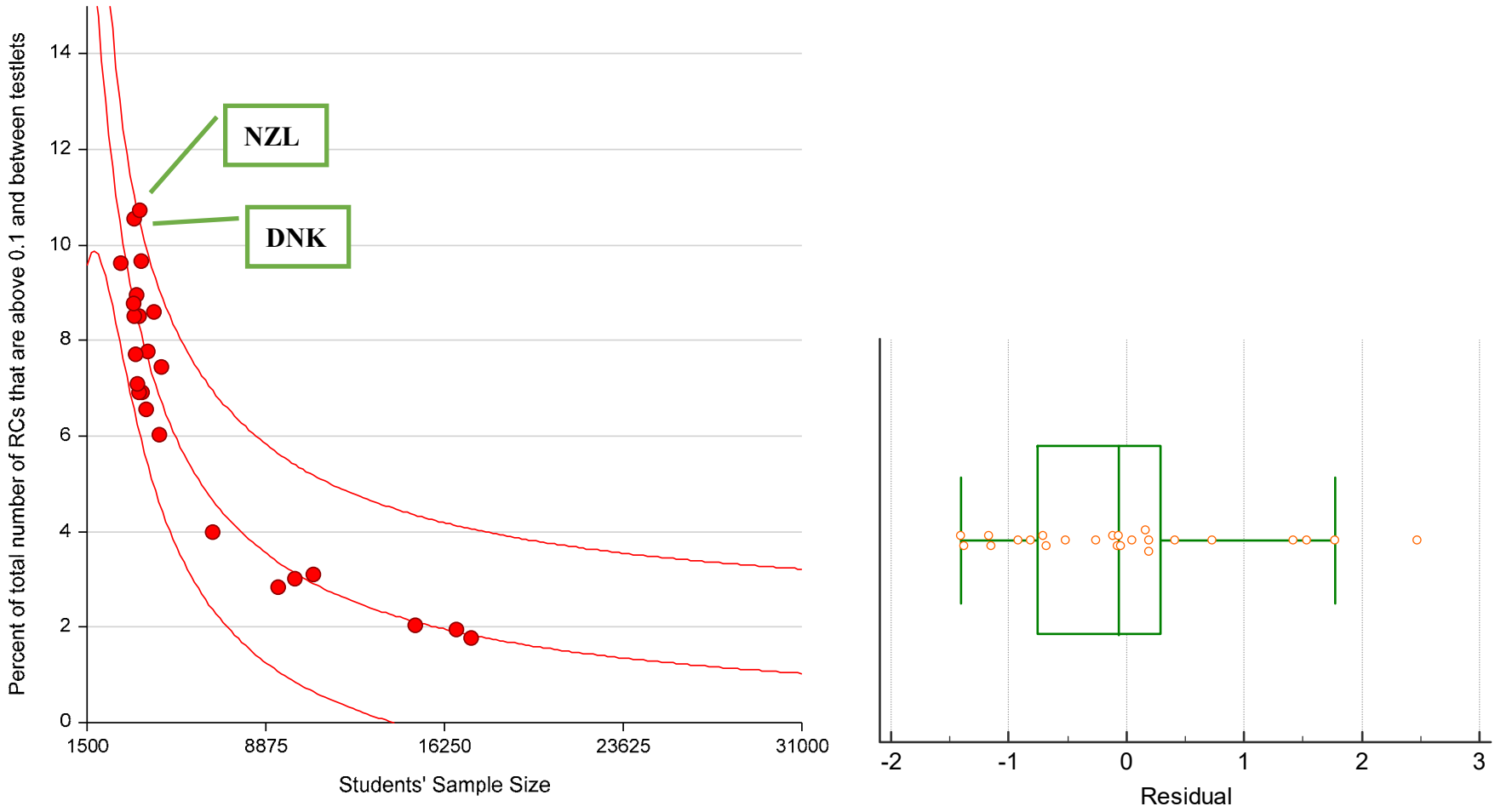


Figure 6.2.5 Reciprocal function and its prediction limits fitted to picture association between student sample size and prevalence of RCs along with a box-plot of prediction residuals (PISA 2006 / RCs that are above 0.1 / Pairs of items from different testlets / Mathematics)

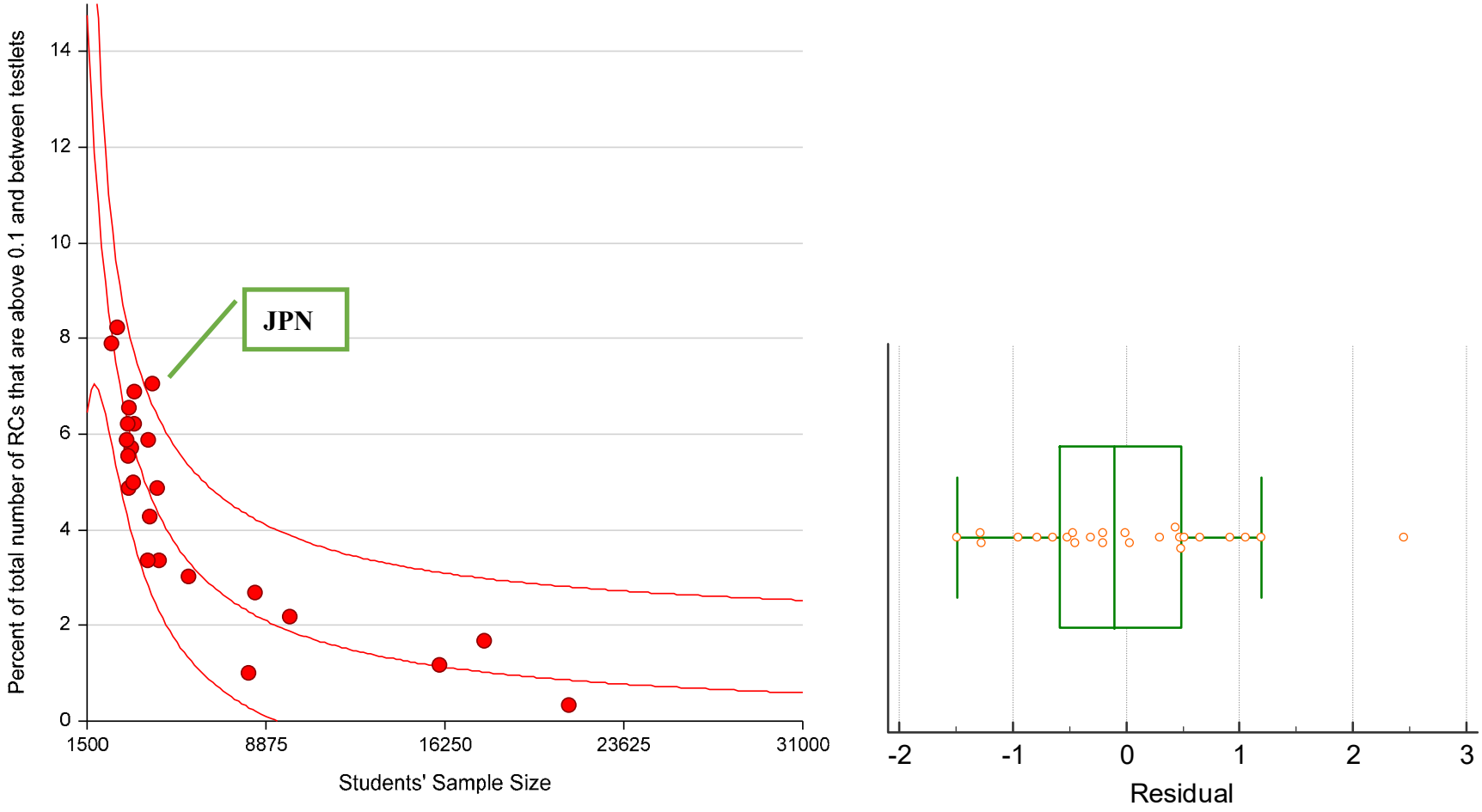


Figure 6.2.6 Reciprocal function and its prediction limits fitted to picture association between student sample size and prevalence of RCs along with a box-plot of prediction residuals (PISA 2009 / RCs that are above 0.1 / Pairs of items from different testlets / Mathematics)

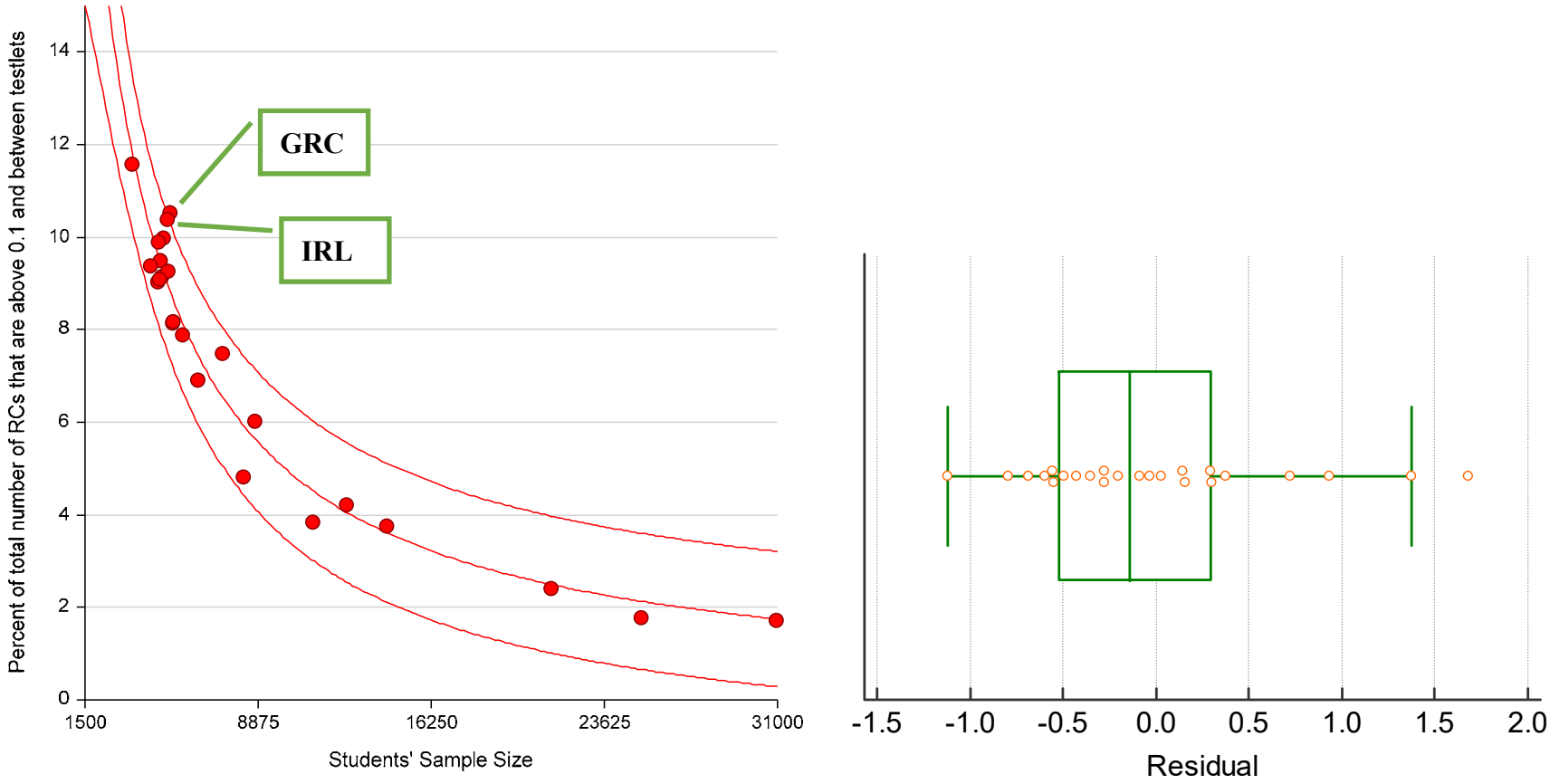
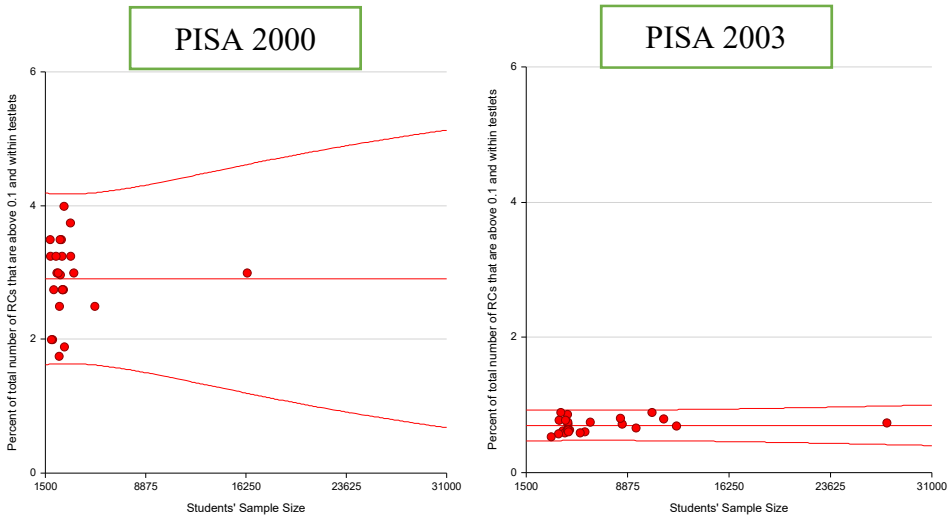


Figure 6.2.7 Reciprocal function and its prediction limits fitted to picture association between student sample size and prevalence of RCs along with a box-plot of prediction residuals (PISA 2012 / RCs that are above 0.1 / Pairs of items from different testlets / Mathematics)

Table 6.2.1 reports three different techniques for accessing outliers. Given that the number of 24 countries involved is small, a liberal p=0.1 level for Grubbs' Single-Outlier Level Test (Alternative Hypothesis: One-Sided vs Maximum) has been used. The small sample size also can explain a lack of consistency in flagging outliers between different procedures. According to the three separate methods, the level of positive between-testlets LID was outlyingly high in Japan for PISA 2009 after controlling for students sample size. Japan also revealed high levels of positive LID in PISA 2000, while Korea and Finland did so in PISA 2003. Any explanations as to why this may be the case can be only speculative. Perhaps in these countries students are able to apply some common skills, such as reading graphs more efficiently as compared to students in other countries, leading to increased dependency between mathematical questions from different testlets.

Figure 6.2.2 suggests that the prevalence of RCs exceeding 0.1 when both items come from the same testlets is less impacted by the student's sample size. This is confirmed in Figure 6.2.8 which mimics the previous five figures but merges them into a single graph. Because the R2s for the reciprocal functions fitted in graphs are close to 0, the outliers investigation was undertaken on raw data and is reported in Table 6.2.2.



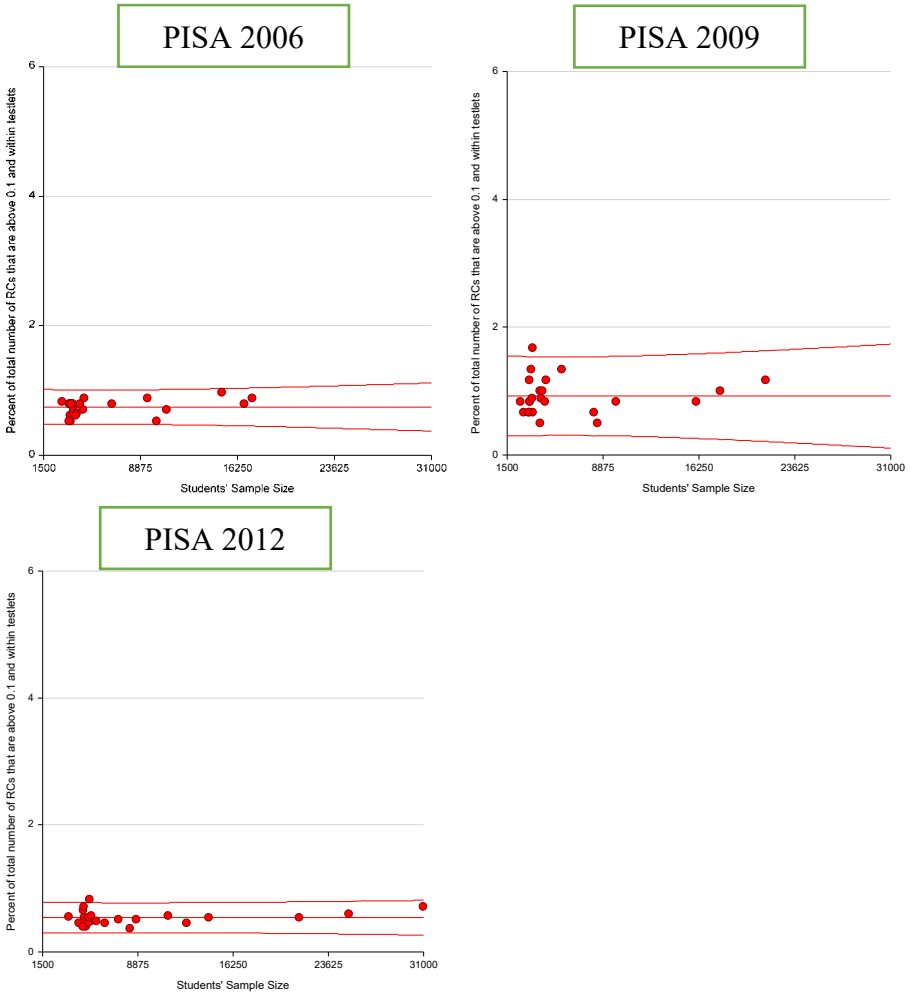


Figure 6.2.8 Reciprocal function and its prediction limits fitted to show the association between students' sample size and prevalence of RCs (PISA 2000,2003,2006,2009 and 2012 / RCs that are above 0.1 / Pairs of items from within the same testlets / Mathematics)

Table 6.2.2 Countries with high levels of within-testlet positive LID in mathematics

	COUNTRY	Value of Possible Outlier	ESD|Z|	Grubbs' Single-Outlier Level Test Prob Level (Alternative Hypothesis: One-Sided vs Maximum)	Conclude Outlier by Rosner's Procedure	Tukey, 1977 - Outside values
PISA 2000	CZECH REP (CZE)	4.0	1.86	0.670	NO	NO
	SPAIN (ESP)	3.7	1.61	1.000	NO	NO
PISA 2003	SPAIN (ESP)	0.9	1.83	0.709	NO	NO
	NORWAY (NOR)	0.9	2.03	0.392	NO	NO
PISA 2006	SPAIN (ESP)	1.0	1.89	0.618	NO	NO
	CANADA (CAN)	0.9	1.33	1.000	NO	NO
	JAPAN (JPN)	0.9	1.42	1.000	NO	NO
	SWITZERLAND (CHE)	0.9	1.53	1.000	NO	NO
PISA 2009	**GREECE (GRE)**	1.7	2.67	**0.044**	**YES**	NO
	GERMANY (DEU)	1.3	1.91	0.551	NO	NO
PISA 2012	**GREECE (GRE)**	0.8	2.71	**0.039**	**YES**	**YES**
	ITALY (ITA)	0.7	2.11	**0.311**	**YES**	NO
	SWEDEN (SWE)	0.7	2.43	**0.099**	**YES**	NO

Grey areas highlight the PISA waves for which mathematics was a targeted domain.

Greece was found to indicate higher levels of within-testlet positive LID in PISA 2009 and PISA 2012 when compared to the other 23 OECD countries. Although there are many causes for LID suggested by Yen (1993) which relate to item formats, passage dependence, or item chaining, these would not be likely to produce increased levels of LID for only one country, as students from most of the countries were exposed to the same set of the mathematical cognitive items. A possible explanation for this result may be student related such as the possibility of increased fatigue or external assistance or interference which were also suggested by Yen (1993) as one of LID drivers. There also may be at play specific teachers' instructional practices or students' learning behaviours that focus on taking advantage of items being clustered in the same testlet. A conclusive causal evidence for any of the LID drivers, proposed in the literature, explaining Greece's results cannot be offered.

While the percentage of within-testlet positive LID indicating pairs of items that are taken out of a total of all RCs can be expected to be small, Figure 6.2.9 reports within-testlet RCs exceeding +0.1 when the total comprises only RCs of item pairs from the within testlets.

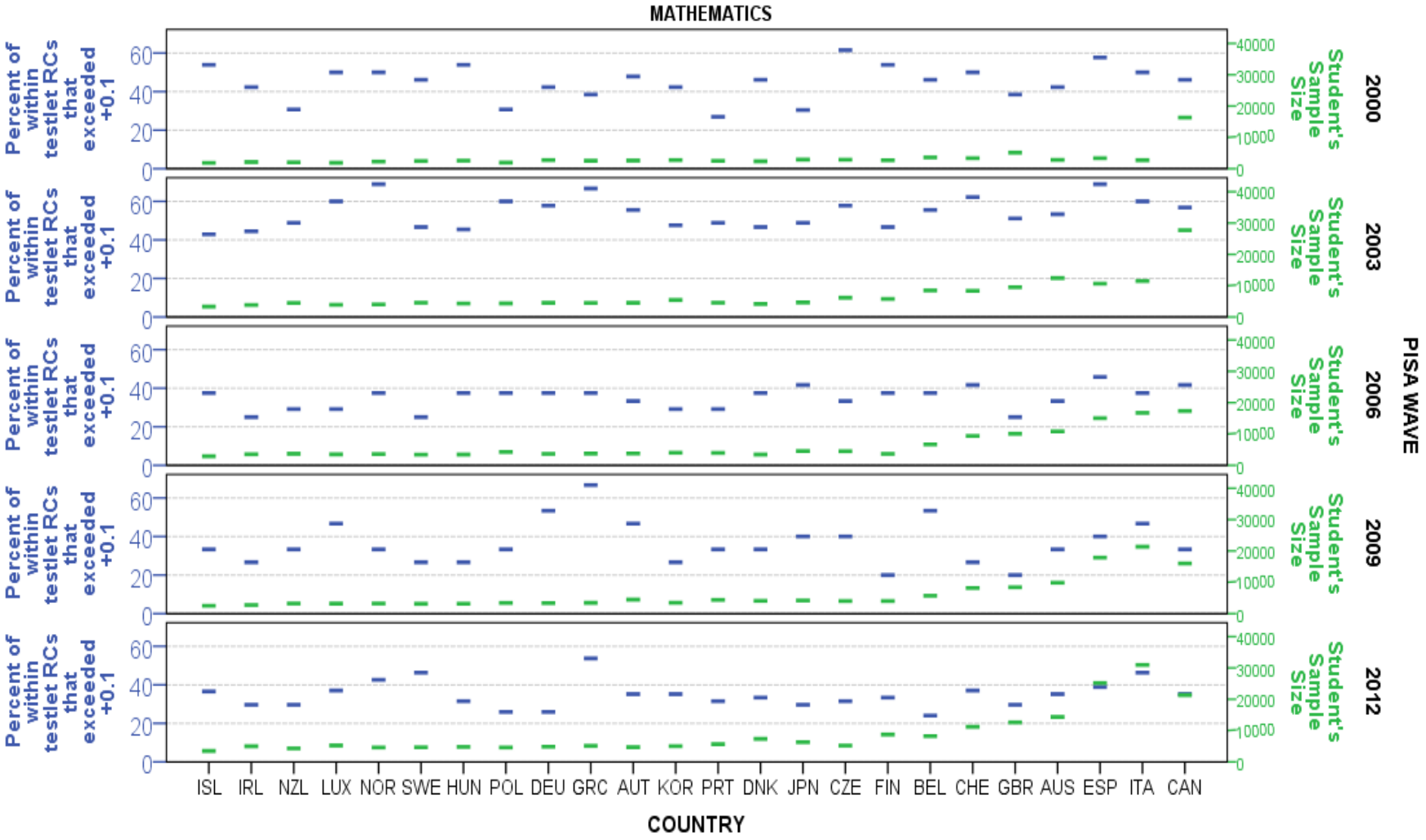


Figure 6.2.9 Dual graph showing the percent of mathematics item pairs with RCs exceeding 0.1 taken out of total of only within-testlet RCs plotted against students sample sizes in mathematics

While Figure 6.2.9 does not add anything new concerning country comparisons, it does offer some cross-wave comparisons that are also aggregated in Figure 6.2.10.

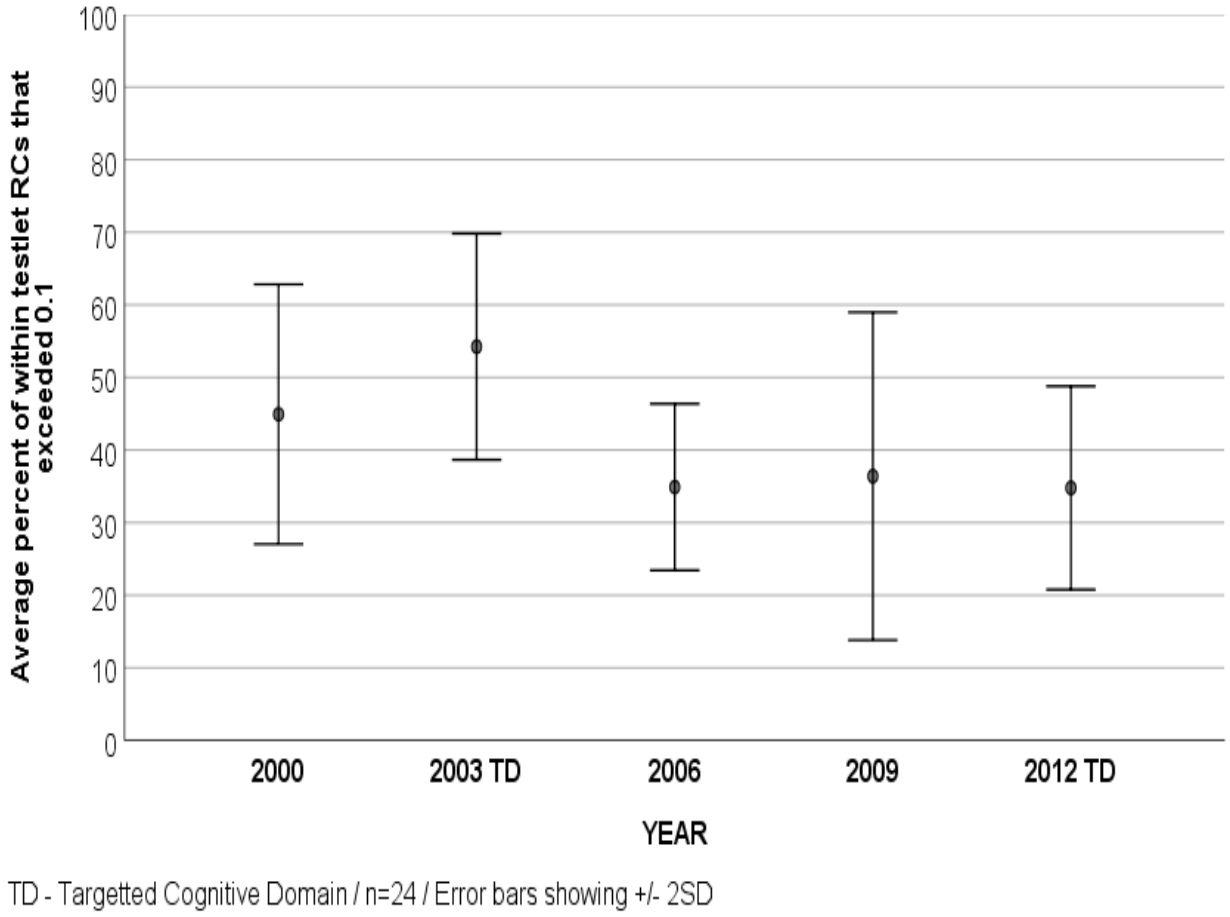


Figure 6.2.10 Average percentage of mathematics item pairs with RCs exceeding 0.1 of total within-testlet RCs obtained from 24 OCED countries.

6.2.2 National calibrations with negative LID in mathematics

While looking at the residual correlations that are lower than -0.1, the sample size related dependency is again present as can be seen in Figure 6.2.11.

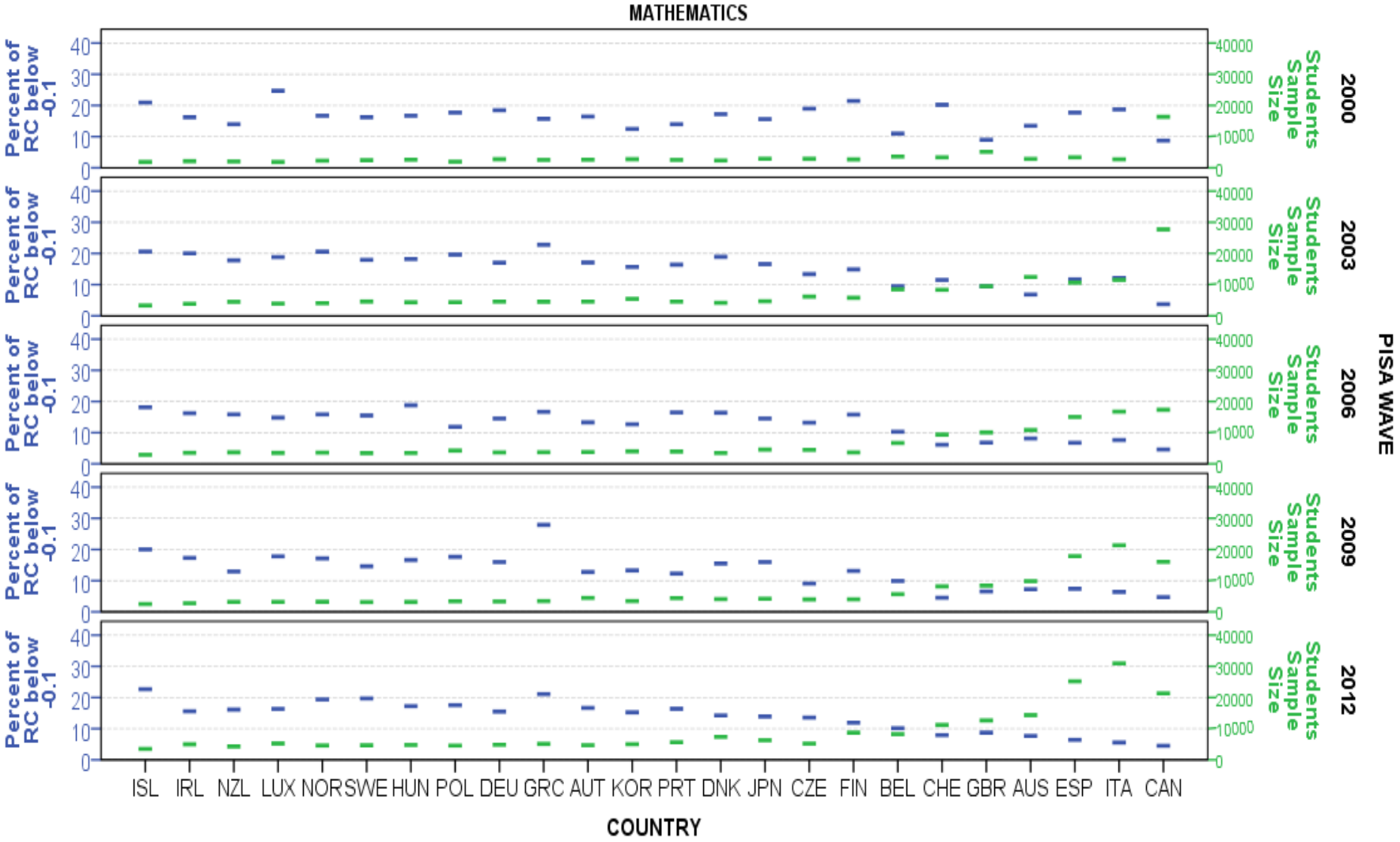


Figure 6.2.11 Dual graph showing the percent of mathematics item pairs with RCs below -0.1 taken out of all RCs against students sample sizes

In the manner similar to the percentage of RCs suggesting positive LID reported above, the outlying countries with high negative LID are investigated using the same methodology. Table 6.2.3 lists countries which may have an unusually high prevalence of negative LID while supporting Figures 6.2.12-6.2.16 graph the fits of reciprocal functions along with Box-and-Whisker plots of prediction residuals.

Table 6.2.3 Countries with high levels of negative LID in mathematics

	R2 FOR Y=1/(A+ BX)	COUNTRY	Value of Possible Outlier	ESD |Z|	Grubbs' Single-Outlier Level Test Prob Level (Alternative Hypothesis: One-Sided vs Maximum)	Conclude Outlier by Rosner's Procedure	Tukey, 1977 - Outside values
PISA 2000	0.38	LUXEMBOURG (LUX)	5.9	1.95	0.526	NO	NO
		SWITZERLAND (CHE)	5.0	1.88	0.605	NO	NO
PISA 2003	0.89	**GREECE (GRC)**	4.8	3.09	**0.006**	**YES**	**YES**
		ITALY (ITA)	3.1	2.76	**0.028**	**YES**	**YES**
		SPAIN (ESP)	2.1	2.52	**0.069**	**YES**	**YES**
PISA 2006	0.89	HUNGARY (HUN)	2.8	1.97	0.498	NO	NO
		ITALY (ITA)	2.6	2.11	0.314	NO	NO
PISA 2009	0.65	**GREECE (GRC)**	12.0	3.63	**0.000**	**YES**	**YES**
		SPAIN (ESP)	3.4	1.79	0.754	NO	NO
PISA 2012	0.88	**GREECE (GRC)**	4.5	2.60	**0.061**	**YES**	NO
		SWEDEN (SWE)	2.1	1.51	1.000	NO	NO

Grey areas highlight the PISA waves for which mathematics was a targeted domain.

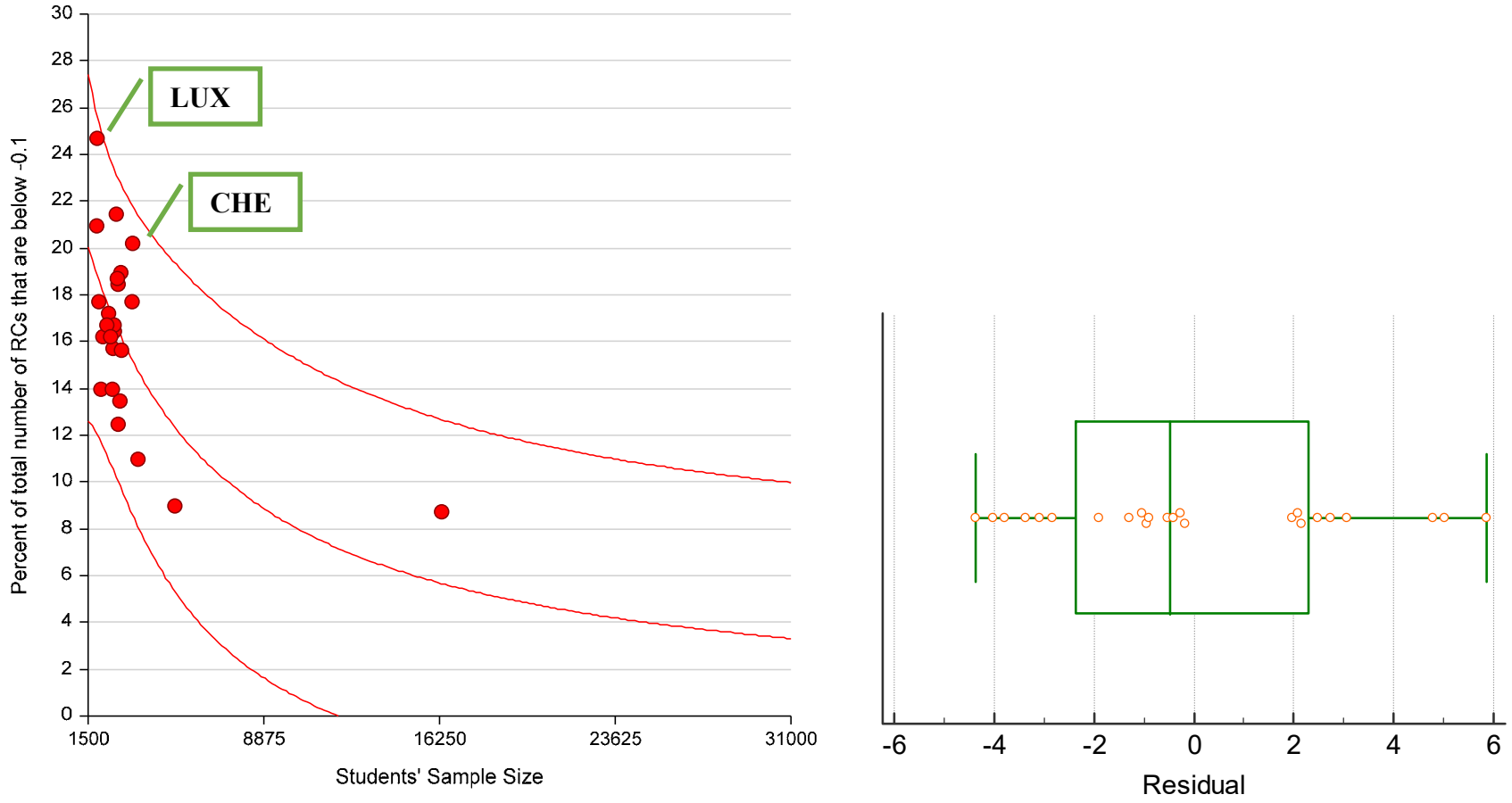


Figure 6.2.12 Reciprocal function and its prediction limits fitted to picture association between student sample size and prevalence of RCs along with a box-plot of prediction residuals (PISA 2000 / RCs that are below -0.1 / Mathematics)

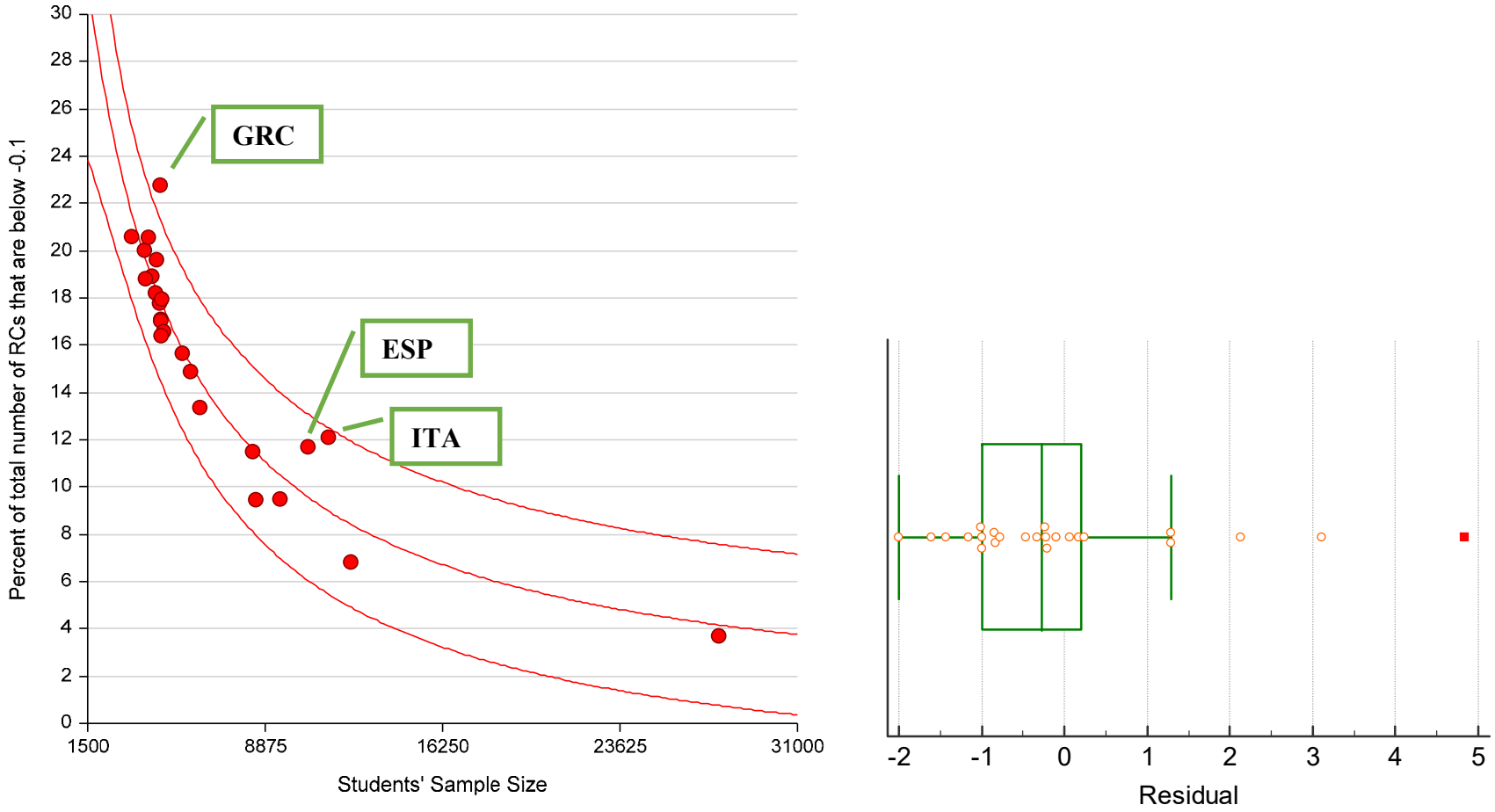


Figure 6.2.13 Reciprocal function and its prediction limits fitted to picture association between student sample size and prevalence of RCs along with a box-plot of prediction residuals (PISA 2003 / RCs that are below -0.1 / Mathematics)

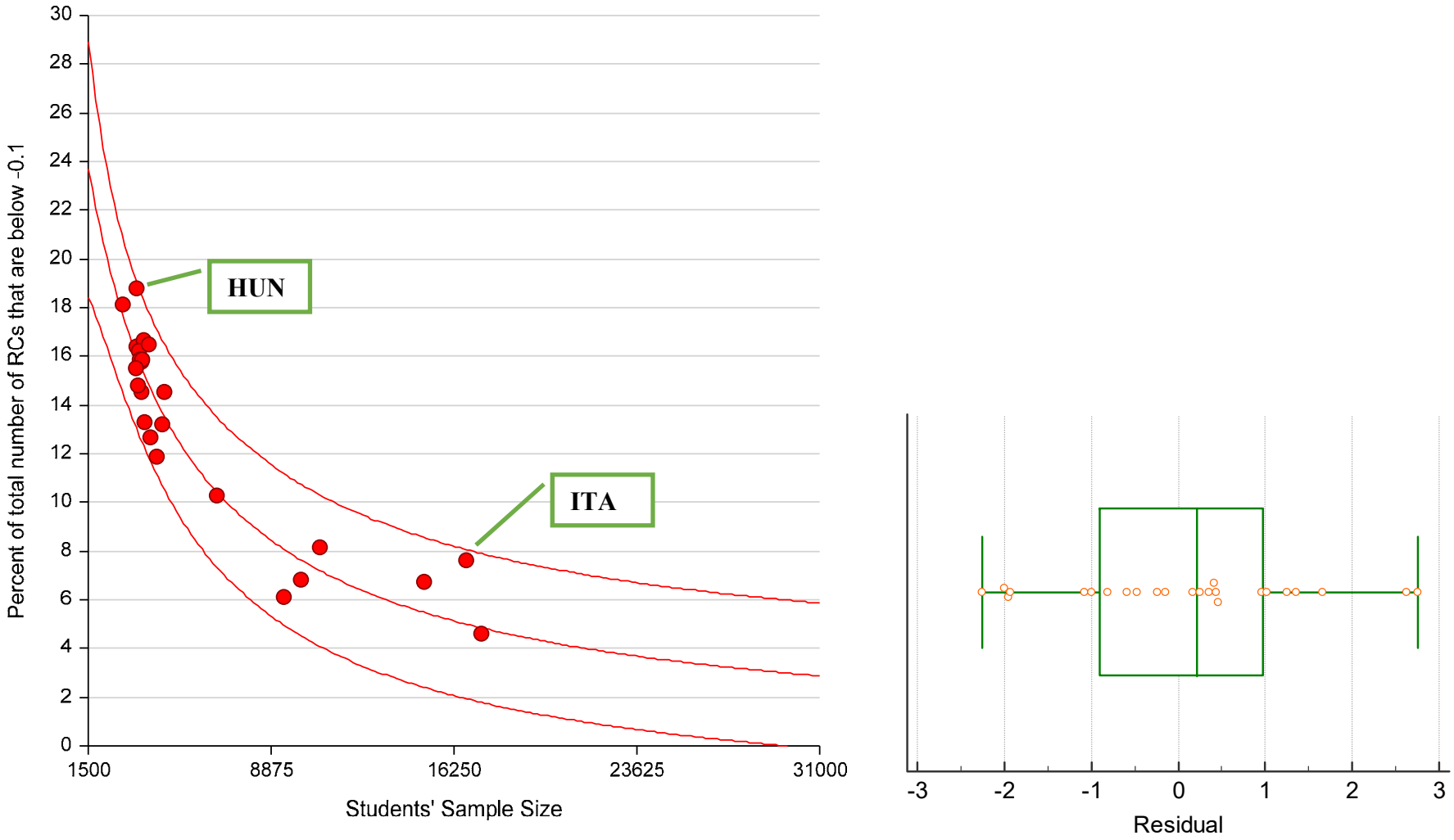


Figure 6.2.14 Reciprocal function and its prediction limits fitted to picture association between student sample size and prevalence of RCs along with a box-plot of prediction residuals (PISA 2006 / RCs that are below -0.1 / Mathematics)

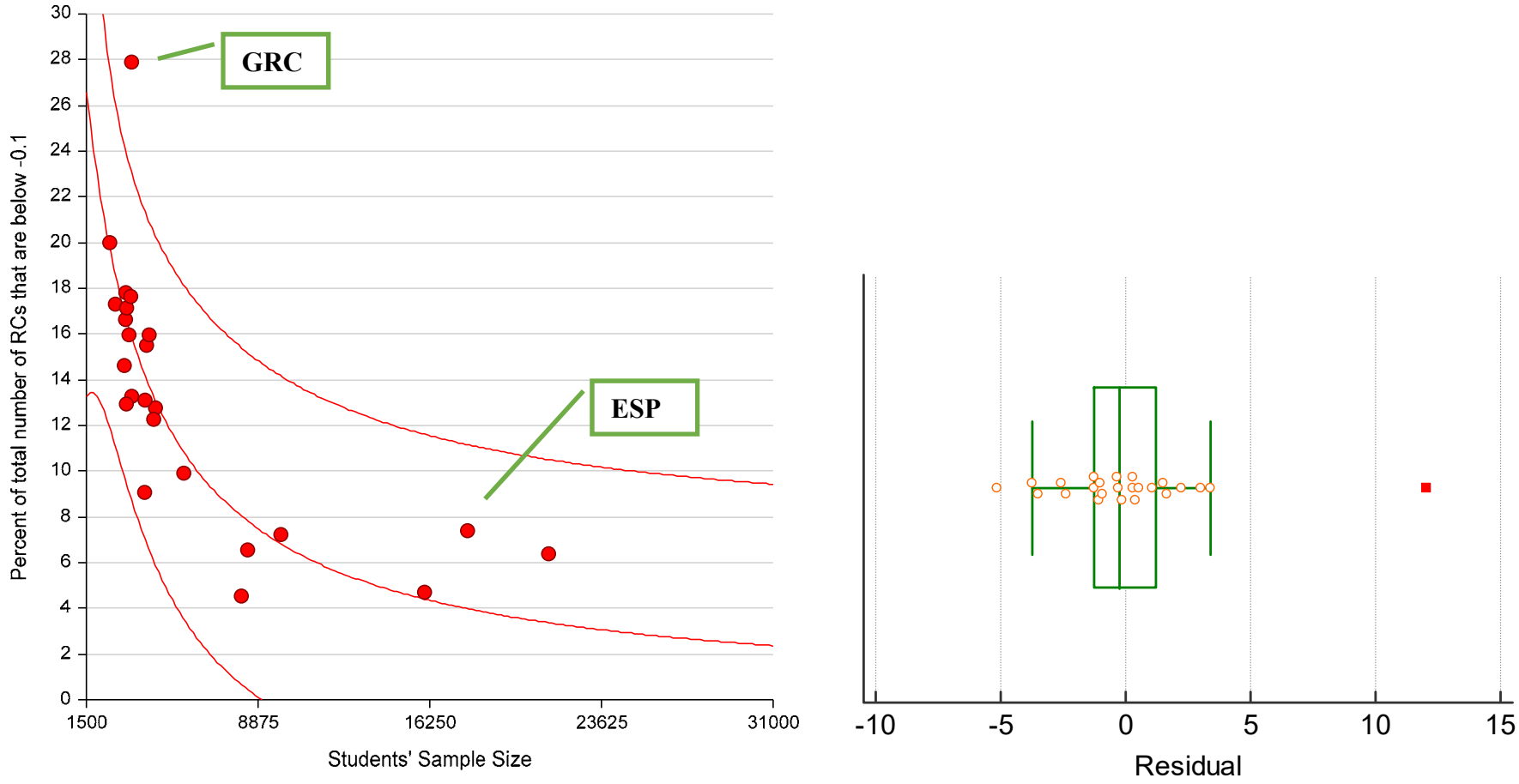


Figure 6.2.15 Reciprocal function and its prediction limits fitted to picture association between student sample size and prevalence of RCs along with a box-plot of prediction residuals (PISA 2009 / RCs that are below -0.1 / Mathematics)

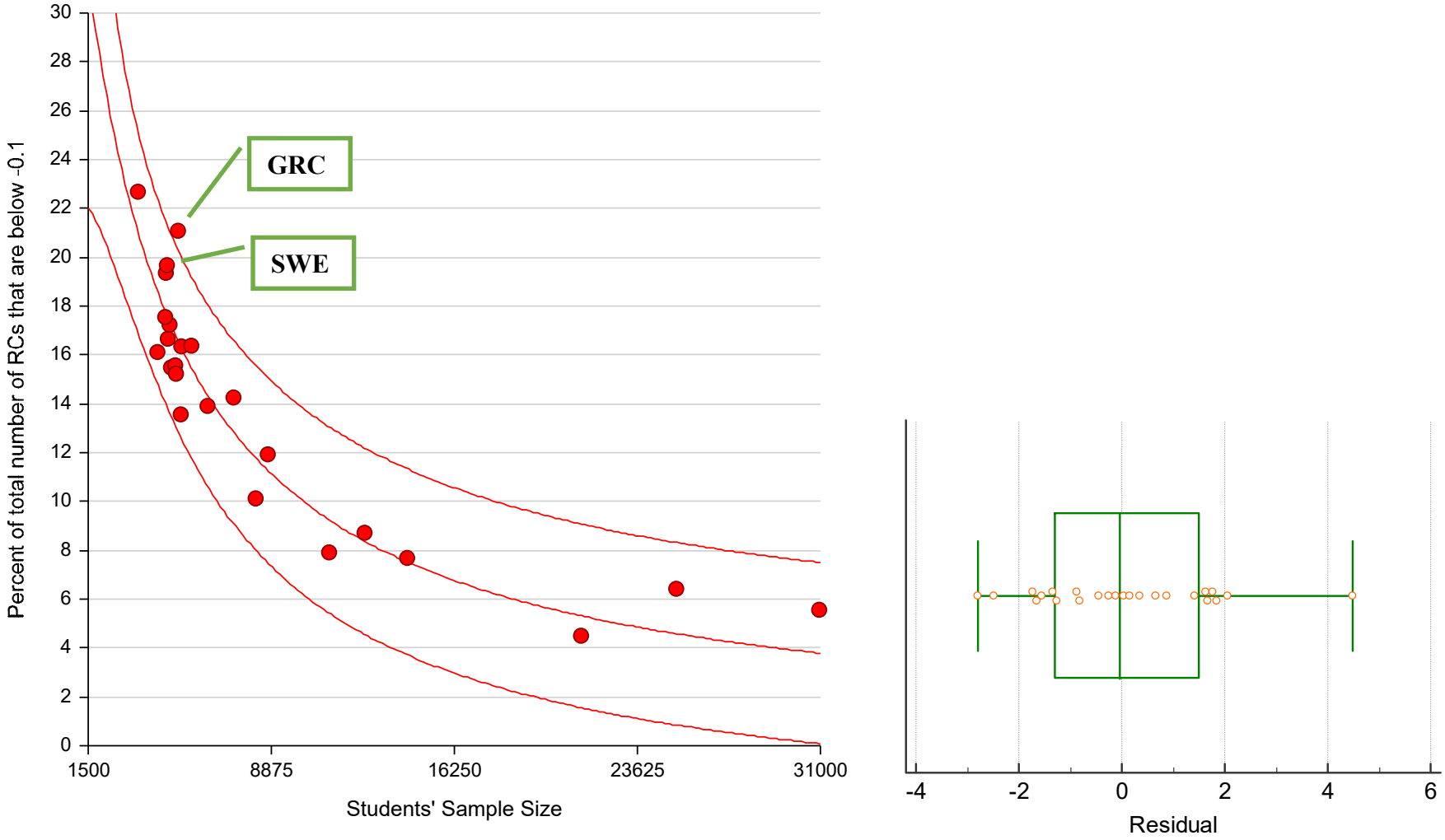


Figure 6.2.16 Reciprocal function and its prediction limits fitted to picture association between student sample size and prevalence of RCs along with a box-plot of prediction residuals (PISA 2012 / RCs that are below -0.1 / Mathematics)

As can be seen in Table 6.2.3, Greece features in three PISA waves as indicating higher levels of negative LID as compared to the majority of other countries. In PISA 2003 also Italy and Spain have a higher proportion of RCs which are lower than -0.1, controlling for students' sample size. As previously mentioned, negative LID can occur when selective time allocation is present (Yen, 1993) or alternatively students do not attempt an item when it looks challenging (as indicative by odds ratio for mathematical items of high difference in Table 5.4.4). Given that all the countries that were highlighted to be outliers in Table 6.2.3 are at the lower achievement end of OECD countries, it could be perhaps expected that their students may apply some selection due to time constraints or perceived difficulties that would result in negative LID.

6.2.3 National calibrations with positive LID in reading

The organisation of this section is consistent with the order of figures and tables introduced in the previous subchapter, mathematics. Figure 6.2.17 presents a dual graph with the percent prevalence of RCs exceeding a cut point of 0.1 (highlighted in blue) against the plot of students' sample sizes (highlighted in green) used in all 120 CFA estimations contribution to the graph. Figure 6.2.18 replicates the results showing positive LID but separates them into the light blue section which reports percentages of RCs from different testlets and a dark blue section with the more than 0.1 RCs for items from the same testlets.

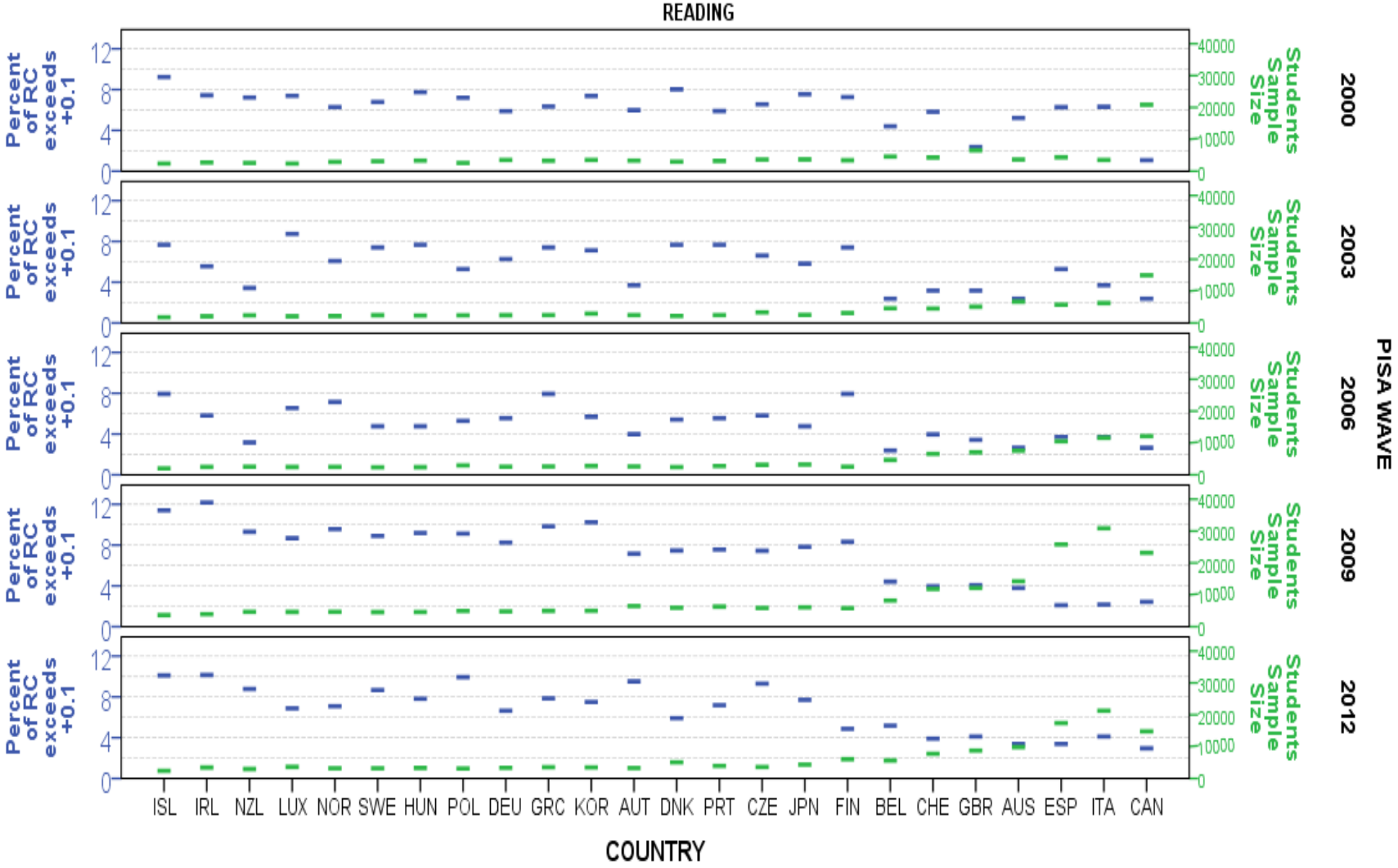


Figure 6.2.17 Dual graph showing the percent of reading item pairs with RCs exceeding 0.1 taken out of all RCs against students' sample sizes

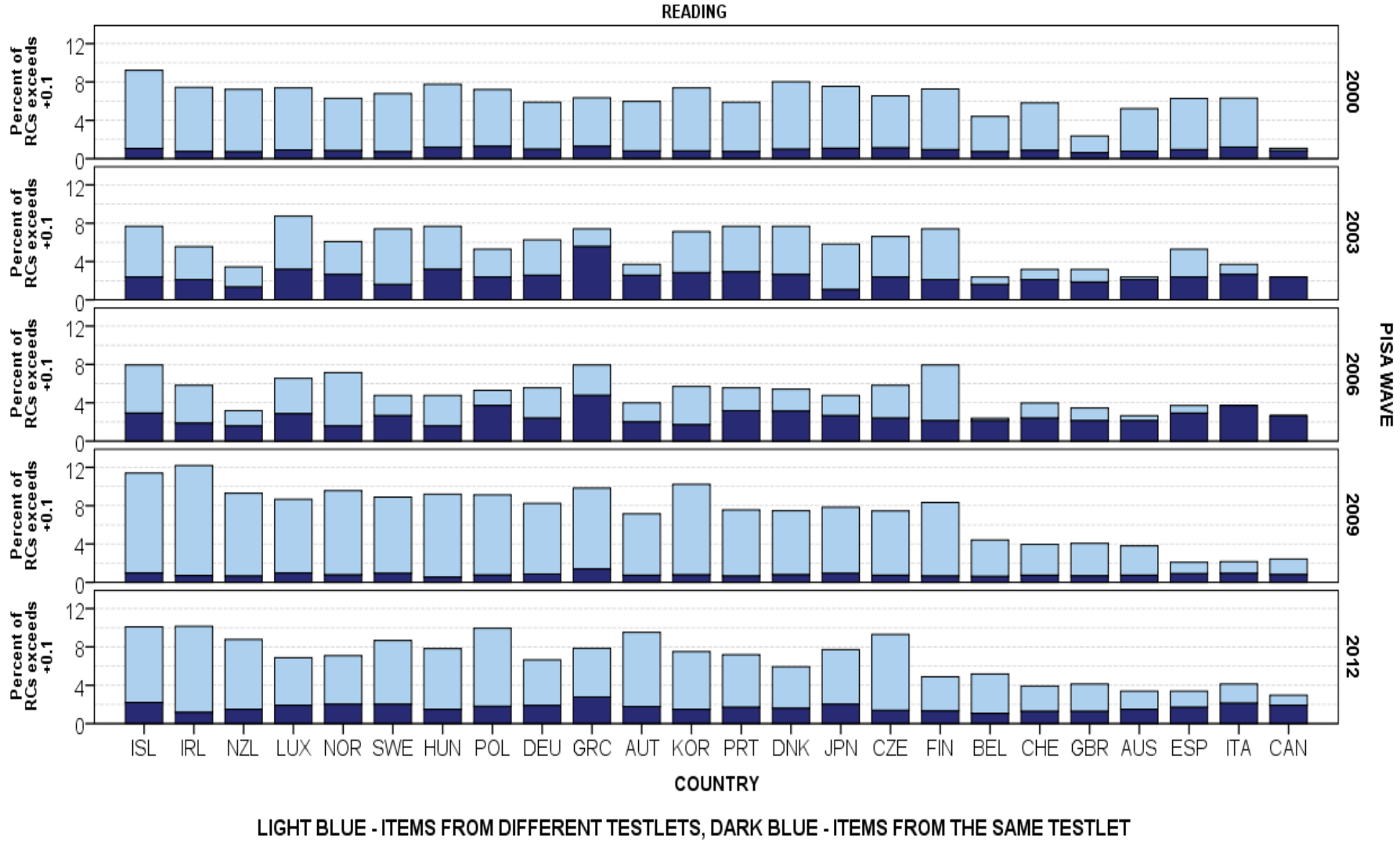


Figure 6.2.18 Percent of reading item pairs with RCs exceeding 0.1 taken out of all RCs separated into components involving item pairs from the same testlets and from different testlets

As was the case in the previous mathematics section, when the students' sample sizes exceed 5000 students the prevalence of positive LID clearly diminishes, putting into question the suitability of a single fixed cut-point for all countries and therefore driving the use of curve fitting in subsequent paragraphs. However, interesting comparisons can be made for countries that used similar size cohorts of students. In PISA 2003, Greece (GRC) and Korea (KOR) reported 7.4% and 7.1% of RCs exceeding the 0.1 cut point when compared to Austria's (AUT) 3.7% estimate[46][47]. Figure 6.2.18 offers an additional view of positive LID, as the estimate for Greece comes largely from within-testlet RCs, yet for Korea, the positive LID is mostly from RCs involving items from different testlets. In the same figure, it is clearly apparent that in PISA 2003 and PISA 2006 the within-testlet percentages are quite high when compared to other years. The distribution of items and testlets used in PISA are represented in Figure 4.3.2 and offers the plausible explanation that after PISA 2000, for which reading was the targeted domain, the same set of only eight testlets (five three-items large, two four-items large and one five-items large) was used in the two subsequent iterations of the study.

Table 6.2.4 below, reports the investigations into which countries may indicate outlying high levels of positive LID involving items from different testlets, and offers a succinct extract from visual results presented in the Figures 6.2.19-6.2.23.

[46] Non-corrected for multiple comparisons 95%CIs for the proportion difference is (0.4%-7%). Epitools were used - http://epitools.ausvet.com.au/content.php?page=z-test-2&p1=0.074&p2=0.037&n1=378&n2=351&Conf=0.05&tails=2&samples=2

[47] Non-corrected for multiple comparisons 95%CIs for the proportion difference is (0.0006%-6.7%). Epitools were used - http://epitools.ausvet.com.au/content.php?page=z-test-2&p1=0.071&p2=0.037&n1=351&n2=351&Conf=0.05&tails=2&samples=2

Table 6.2.4 Countries with high levels of between-testlets positive LID in reading

	R2 FOR Y=1/(A +BX)	COUNTRY	Value of Possible Outlier	ESD |Z|	Grubbs' Single-Outlier Level Test Prob Level (Alternative Hypothesis: One-Sided vs Maximum)	Conclude Outlier by Rosner's Procedure	Tukey, 1977 - Outside values
PISA 2000	0.78	Japan (JPN)	1.3	1.71	0.969	No	No
		Korea (KOR)	1.2	1.74	0.854	No	No
PISA 2003	0.49	Finland (FIN)	2.3	1.81	0.757	No	No
		Sweden (SWE)	2.0	1.75	0.825	No	No
PISA 2006	0.59	**Finland (FIN)**	2.6	2.43	**0.116**	**Yes**	**Yes**
		Norway (NOR)	2.2	2.54	**0.071**	**Yes**	No
PISA 2009	0.95	**Korea (KOR)**	1.4	2.23	0.228	No	**Yes**
		Ireland (IRL)	1.2	2.31	0.166	No	**Yes**
PISA 2012	0.8	**Ireland (IRL)**	2.7	2.59	**0.061**	**Yes**	**Yes**
		Czech Rep (CZE)	1.8	2.25	0.202	No	**Yes**

Grey areas highlight the PISA waves for which reading was a targeted domain.

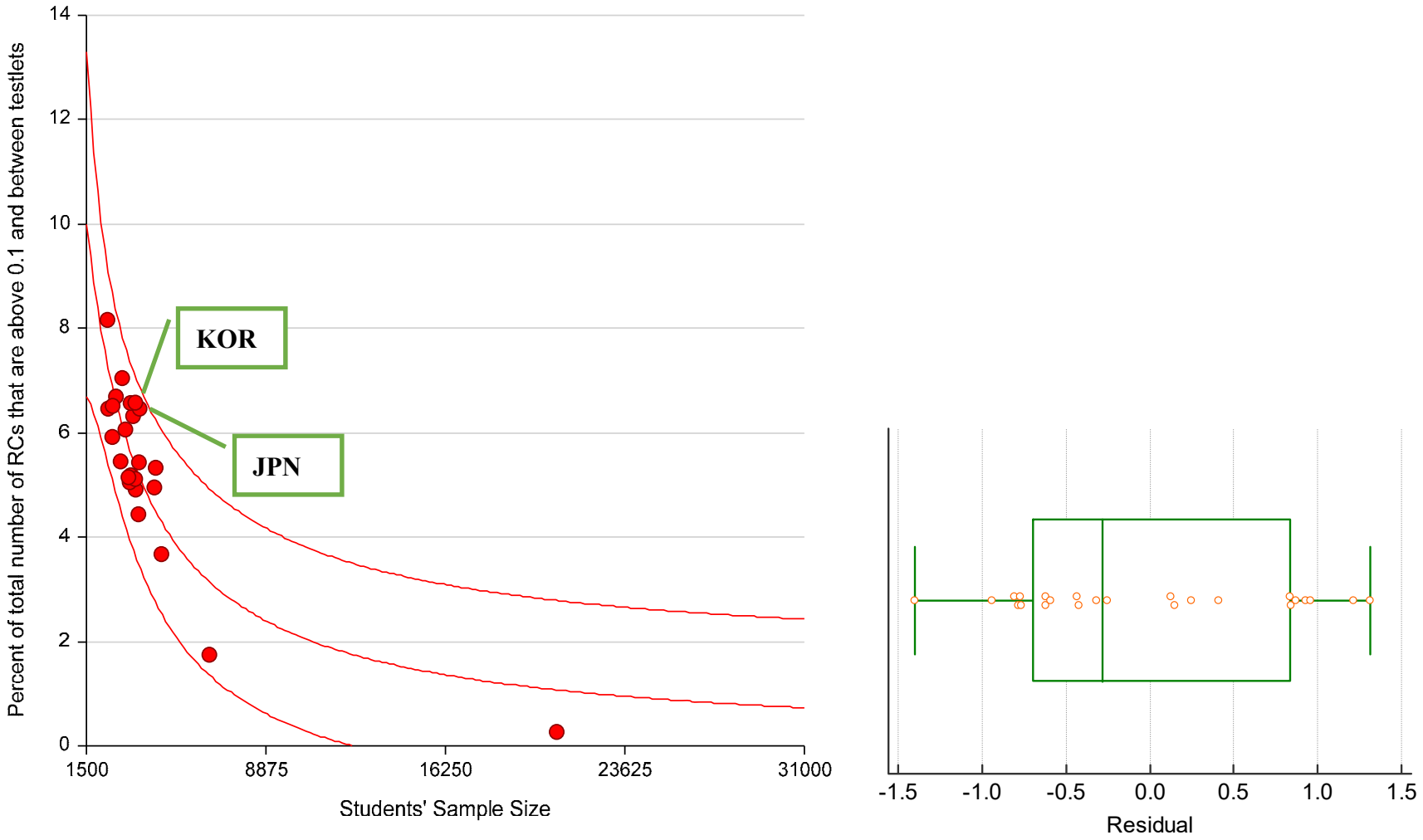


Figure 6.2.19 Reciprocal function and its prediction limits fitted to picture association between student sample size and prevalence of RCs along with a box-plot of prediction residuals (PISA 2000 / RCs that are above 0.1 / Pairs of items from different testlets / Reading)

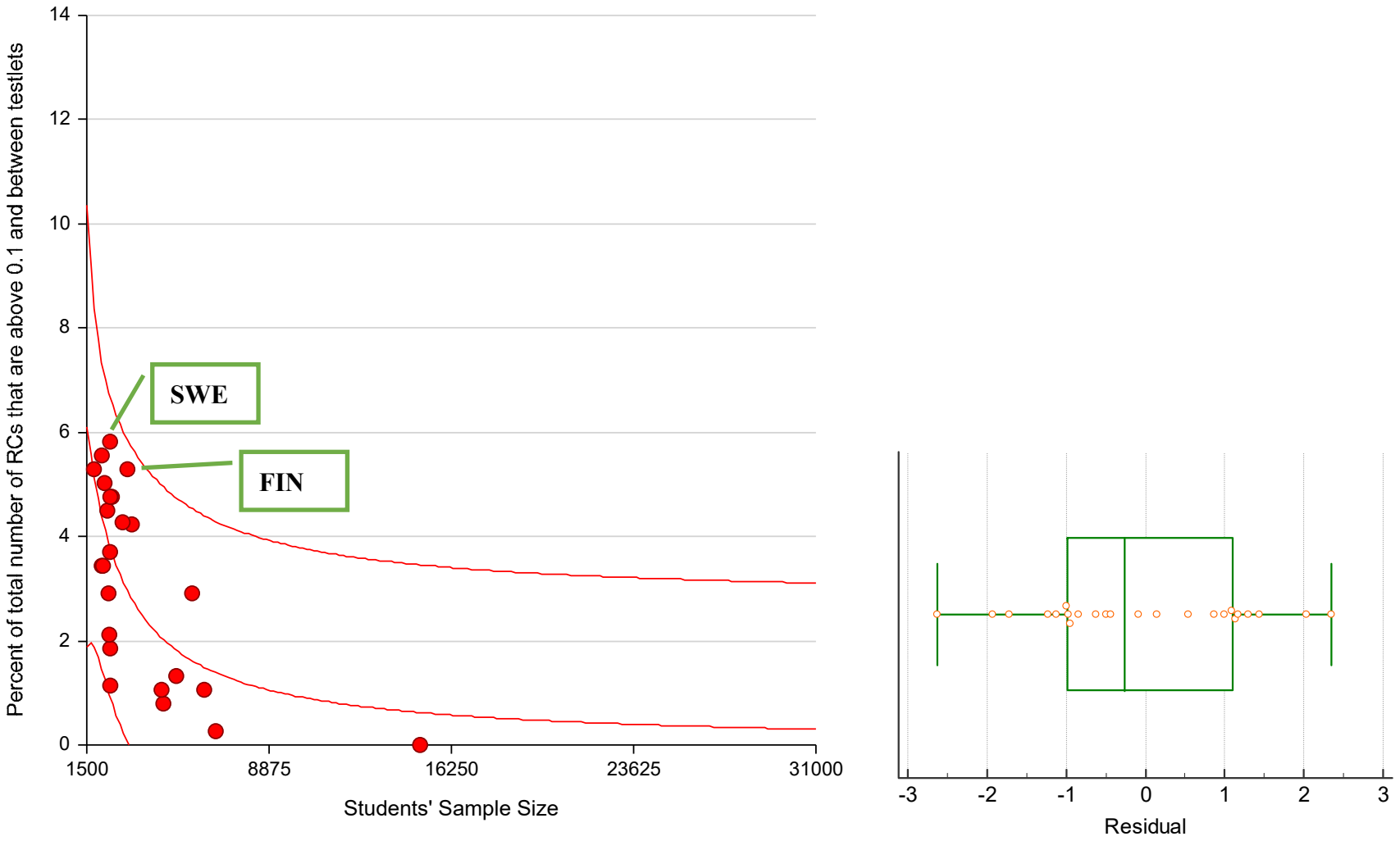


Figure 6.2.20 Reciprocal function and its prediction limits fitted to picture association between student sample size and prevalence of RCs along with a box-plot of prediction residuals (PISA 2003 / RCs that are above 0.1 / Pairs of items from different testlets / Reading)

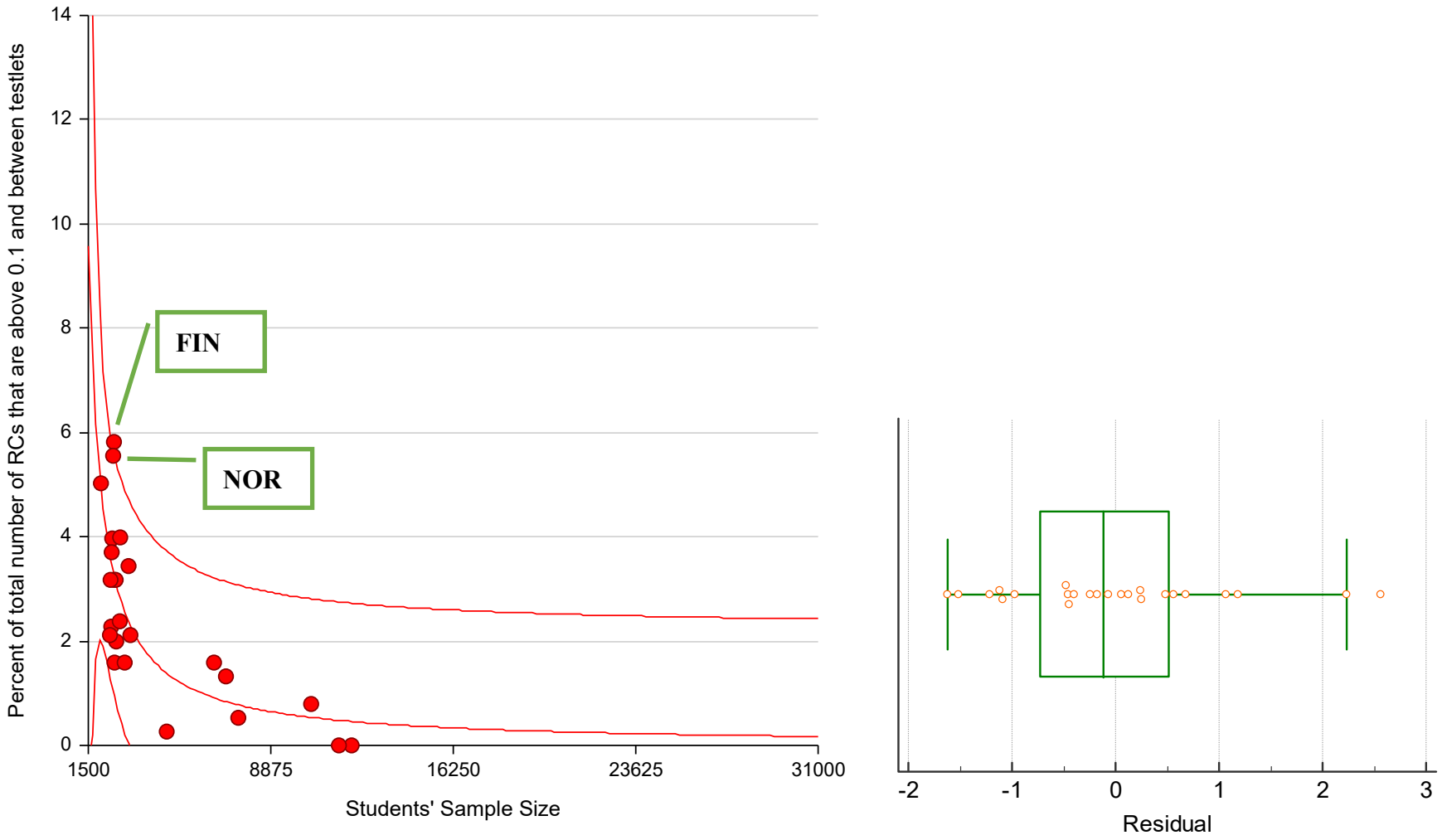


Figure 6.2.21 Reciprocal function and its prediction limits fitted to picture association between student sample size and prevalence of RCs along with a box-plot of prediction residuals (PISA 2006 / RCs that are above 0.1 / Pairs of items from different testlets / Reading)

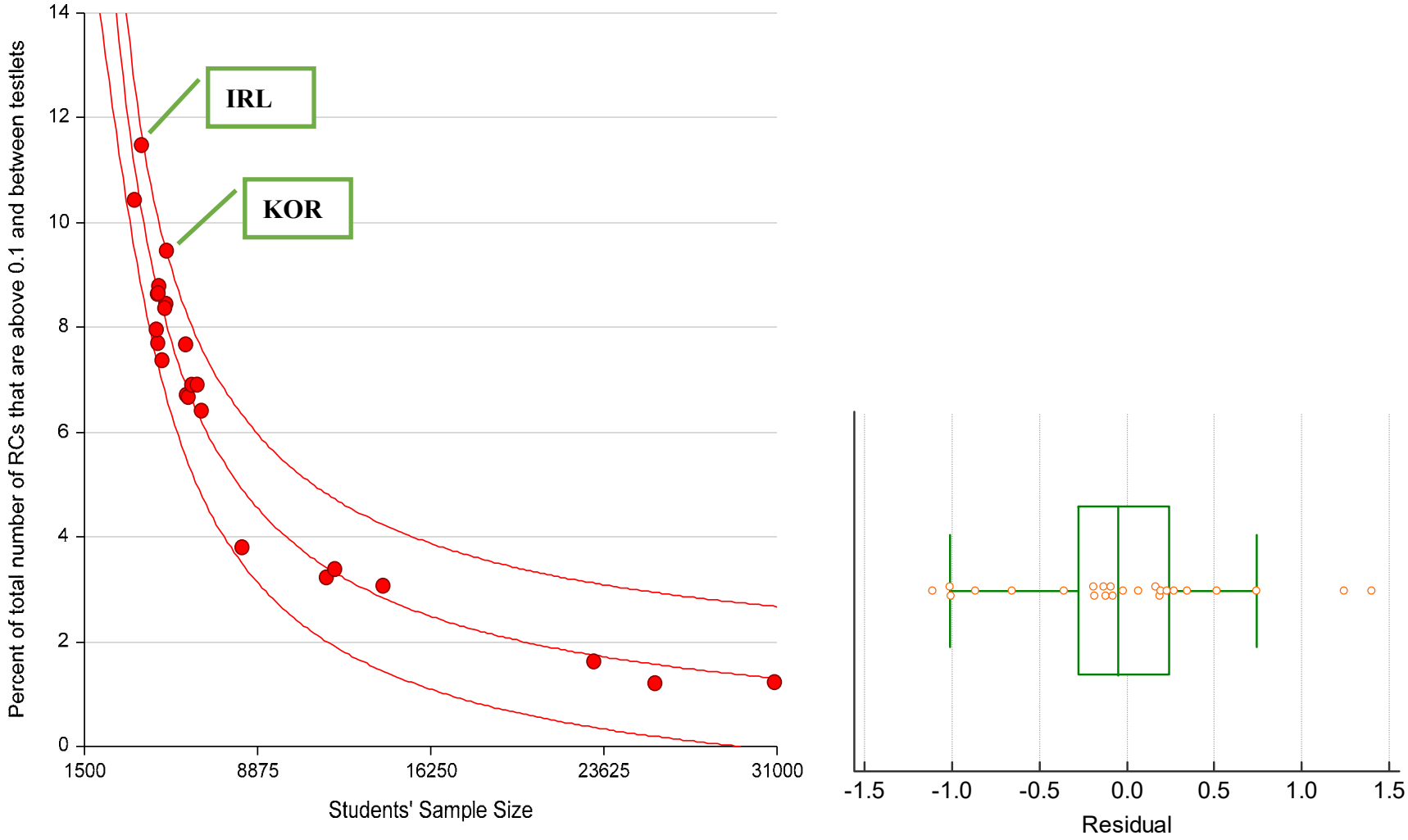


Figure 6.2.22 Reciprocal function and its prediction limits fitted to picture association between student sample size and prevalence of RCs along with a box-plot of prediction residuals (PISA 2009 / RCs that are above 0.1 / Pairs of items from different testlets / Reading)

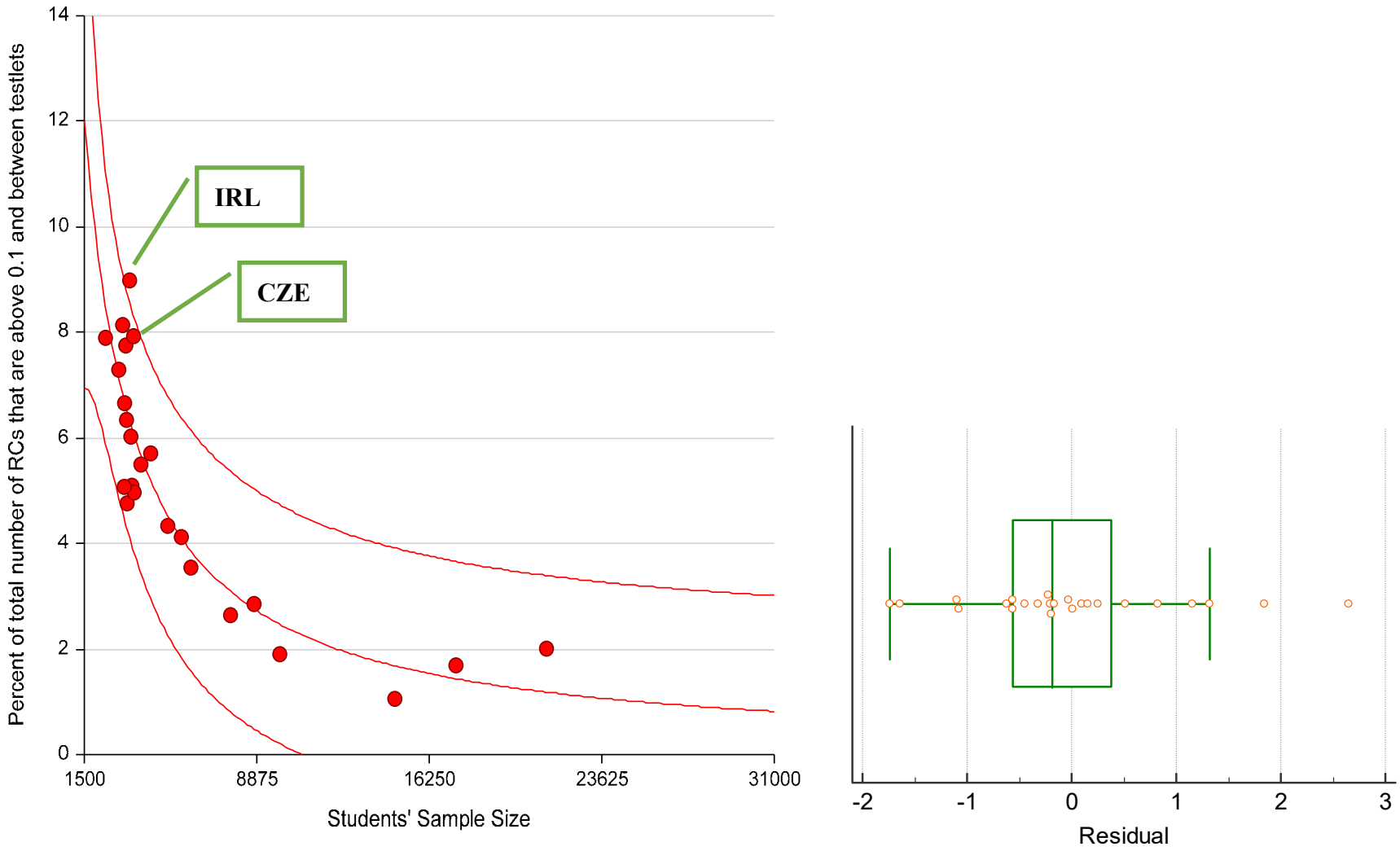
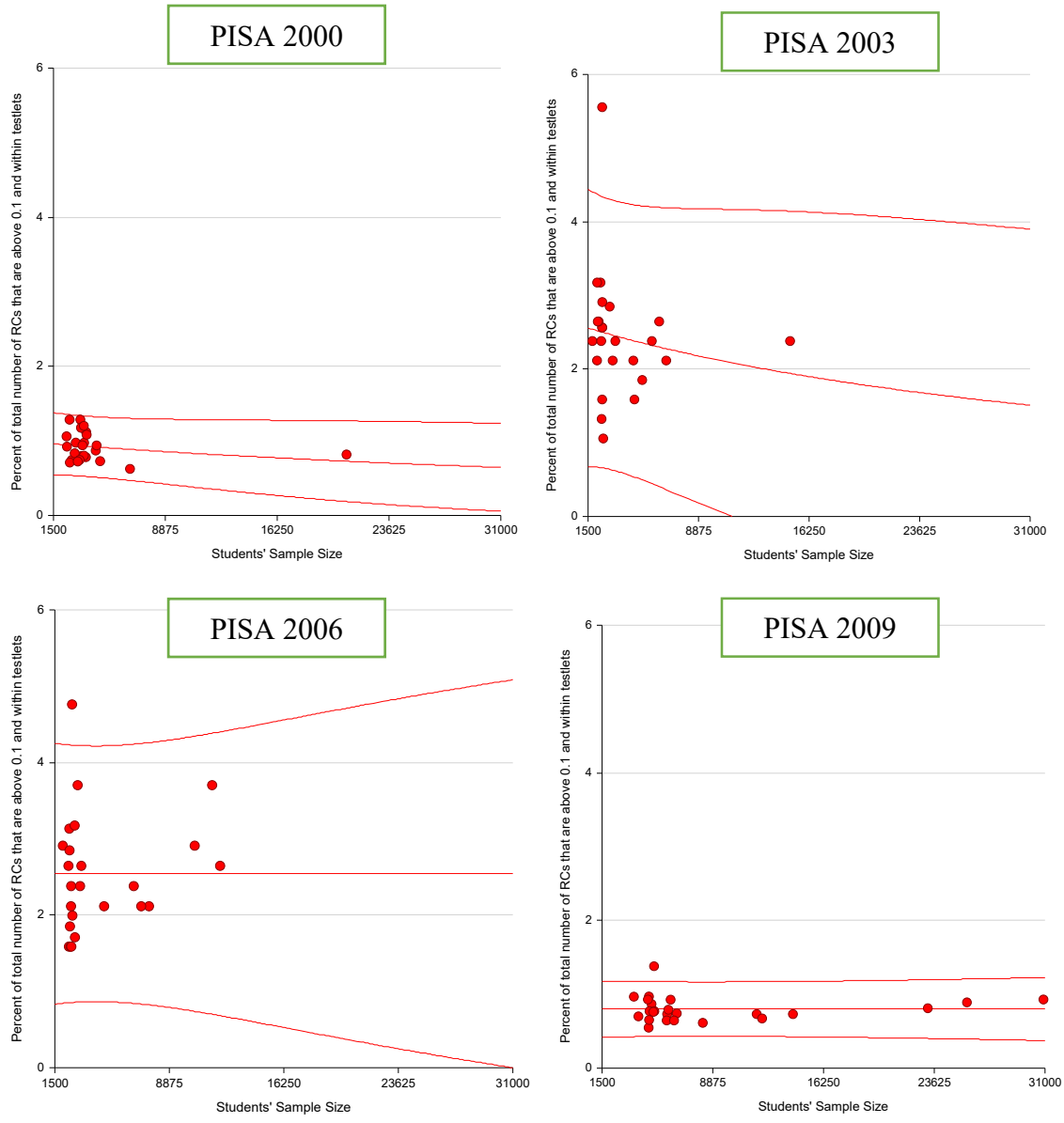


Figure 6.2.23 Reciprocal function and its prediction limits fitted to picture association between student sample size and prevalence of RCs along with a box-plot of prediction residuals (PISA 2012 / RCs that are above 0.1 / Pairs of items from different testlets / Reading)

Table 6.2.4 suggests that for PISA 2006, Finland (FIN) and also possibly Norway (NOR) have higher than other countries' levels of between-testlets positive LID. However in PISA 2009, when reading became again the major targeted cognitive domain and a large number of new testlets and items were given to students, Korea (KOR) and Ireland (IRL) are identified as outlying in regard to between-testlet positive LID. Ireland also features as an outlier in PISA 2012. It is possible that in the countries reported above; students have taken advantage of similar items' formats more efficiently when compared to other nations.

Looking from the perspective of positive LID within testlets, similar to mathematics, the association with students' sample sizes was negligible, as can be seen in Figure 6.2.24. Therefore, Table 6.2.5 reports the outliers on raw data, not the reciprocal function prediction residuals.



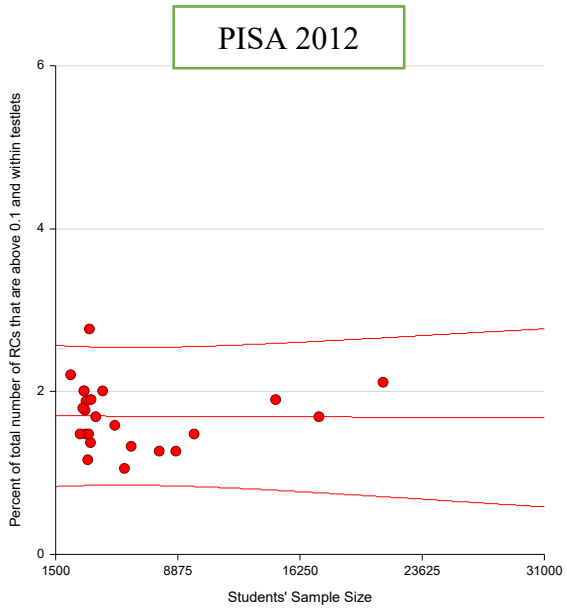


Figure 6.2.24 Reciprocal function and its prediction limits fitted to show the association between students' sample size and prevalence of RCs (PISA 2000,2003,2006,2009 and 2012 / RCs that are above 0.1 / Pairs of items from within the same testlets / Reading)

Table 6.2.5 Countries with high levels of within-testlet positive LID in reading

	COUNTRY	Value of Possible Outlier	ESD|Z|	Grubbs' Single-Outlier Level Test Prob Level (Alternative Hypothesis: One-Sided vs Maximum)	Conclude Outlier by Rosner's Procedure	Tukey, 1977 - Outside values
PISA 2000	Greece (GRC)	1.3	1.90	0.603	No	No
	Poland (POL)	1.3	2.11	0.308	No	No
PISA 2003	**Greece (GRC)**	5.6	3.66	0.000	**Yes**	**Yes**
	Luxembourg (LUX)	3.2	1.59	1.000	No	No
PISA 2006	**Greece (GRC)**	4.8	2.87	**0.018**	**Yes**	**Yes**
	Italy (ITA)	3.7	2.01	0.423	No	No
PISA 2009	**Greece (GRC)**	1.4	3.40	**0.001**	**Yes**	**Yes**
	Luxembourg (LUX)	1.0	1.62	1.000	No	No
PISA 2012	**Greece (GRC)**	2.8	2.74	**0.033**	**Yes**	**Yes**
	Iceland (ISL)	2.2	1.72	0.902	No	No

Grey areas highlight the PISA waves for which reading was a targeted domain.

As was the case in the last two waves of PISA, reported in the corresponding section on mathematics (see Table 6.2.2 for reference), Greece once again features in reading, as achieving the highest levels of within-testlet positive LID out of all the 24 investigated OECD countries. No additional explanations, beyond those offered in the sibling section on

mathematics, can be offered to explain Greece's position. Interestingly, Iceland (ISL) and Luxembourg (LUX), while not formally identified as an outlier, are listed in three PISA waves as countries with a higher prevalence of positive intra-testlet LID. Iceland also featured in two PISA waves in Table 6.2.2 reporting mathematical results. Only speculative explanations could be offered as to why these two countries are featured. Although from one wave of the PISA study to another, a different cohort of students took part in the assessment, in small countries such as Iceland and Luxembourg it is likely that exactly the same schools are being sampled for the study. It is possible that the same teachers look after the study implementation from wave to wave and it is a possibility that in preparation for the assessment they offer new cohorts of students useful suggestions regarding the efficient utilisation of the testlets' stimuli.

As an additional perspective on within-testlet positive LID Figure, 6.2.25 represents the percent of within-testlet RCs exceeding 0.1 when the number of all within-testlet pairs of items is considered.

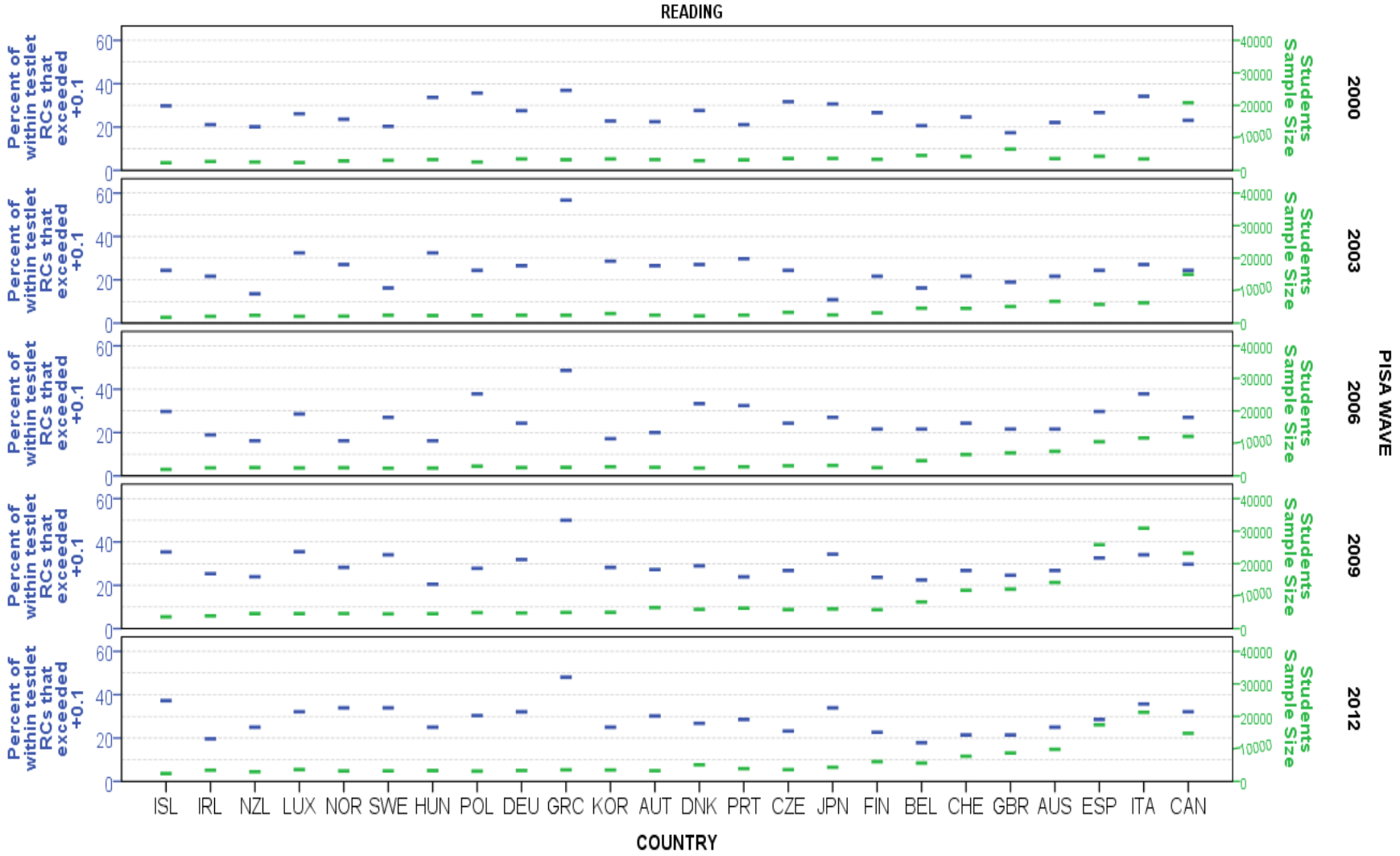


Figure 6.2.25 Dual graph showing the percent of reading item pairs with RCs exceeding 0.1 taken out of total of only within-testlet RCs against students sample sizes

A common reading passage is often presented as an example when local item dependence is discussed. However, as can be seen in Figure 6.2.26, the overall percent of positive LID pairs out of all within-testlet pairs of items is smaller compared to similar results in Figure 6.2.10 representing mathematics.

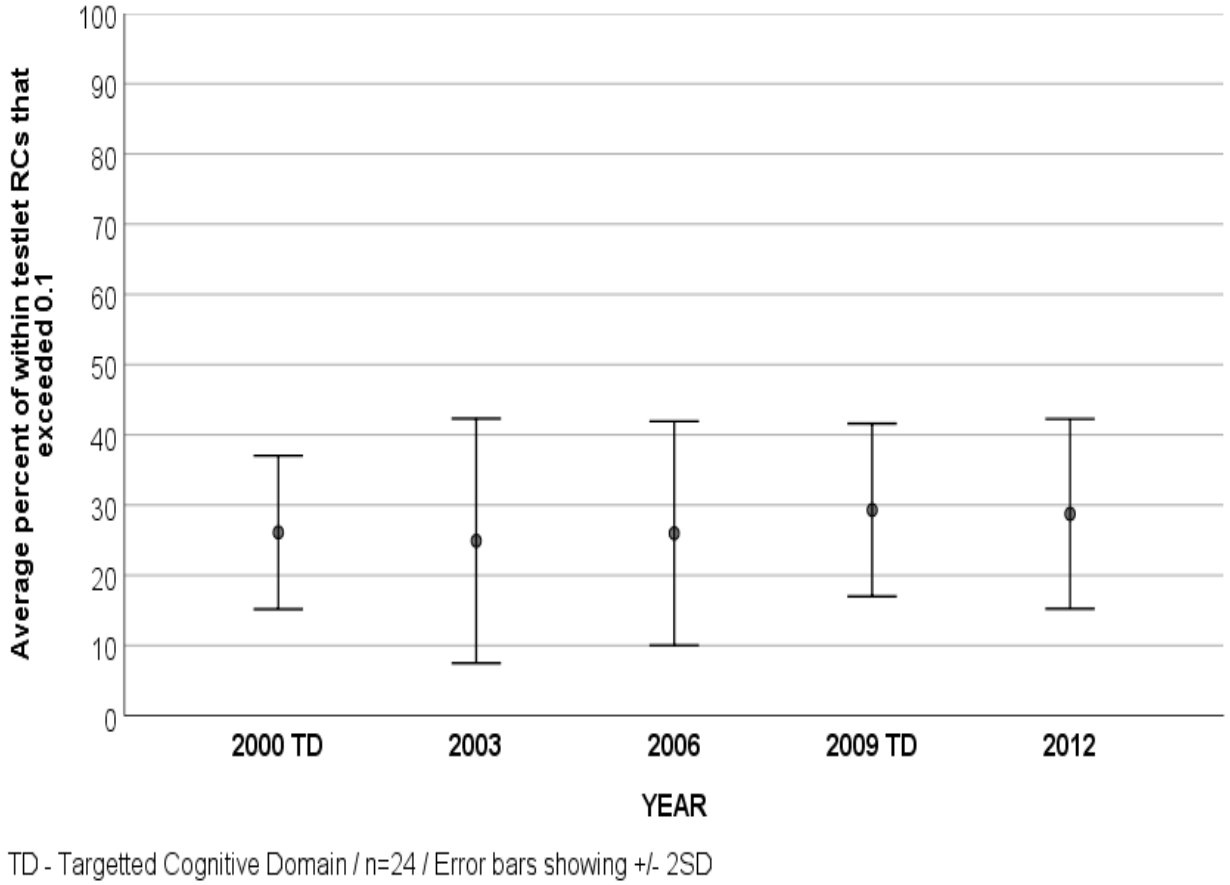


Figure 6.2.26 Average percentage of reading item pairs with RCs exceeding 0.1 of total within-testlet RCs obtained from 24 OCED countries.

6.2.4 National calibrations with negative LID in reading

The proportions of the residual correlation being lower than -0.1 are shown in Figure 6.2.27.

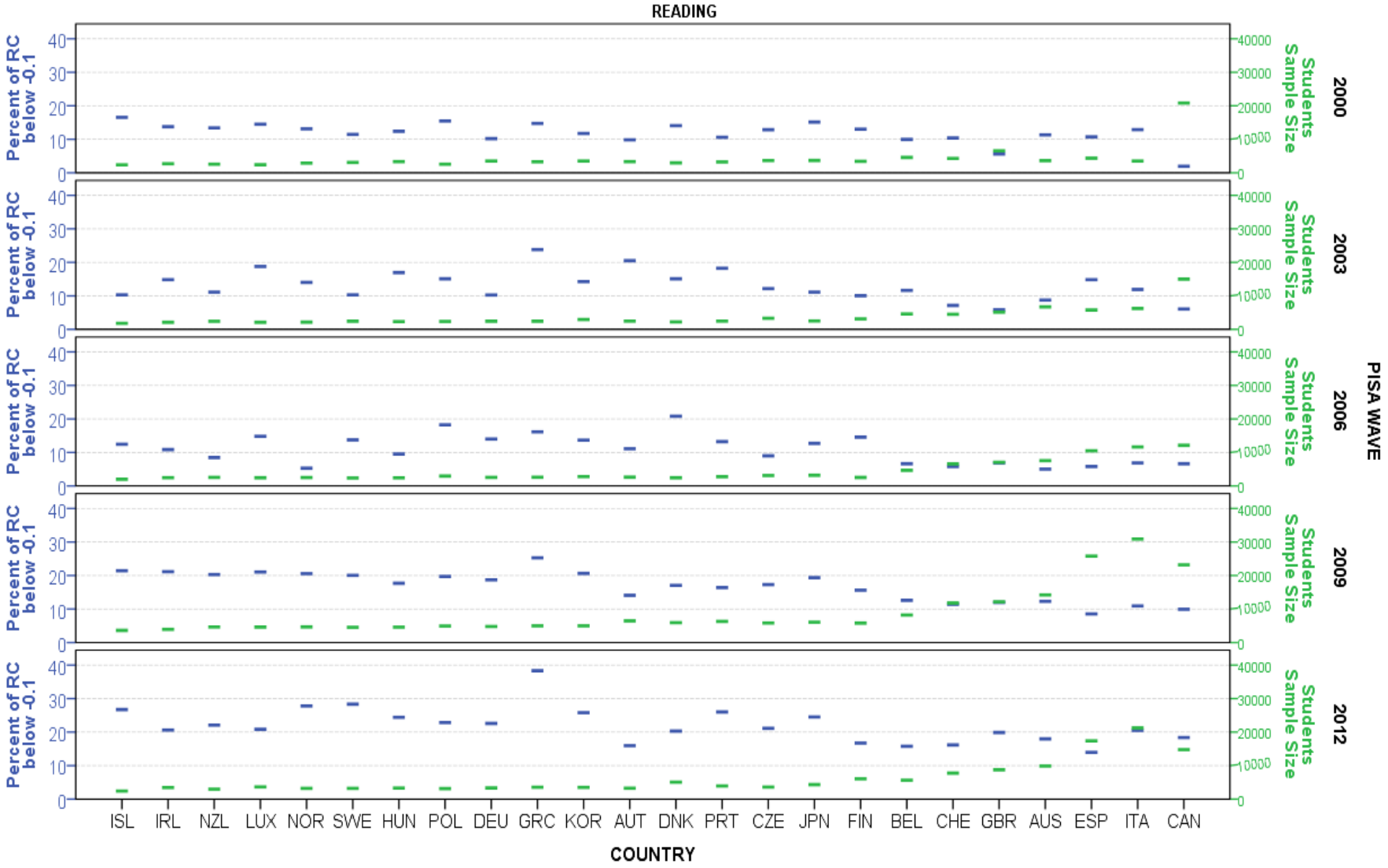


Figure 6.2.27 Dual graph showing the percent of reading item pairs with RCs below -0.1 taken out of all RCs against students' sample size

The relationship between students' sample size and the prevalence of negative LID was less pronounced compared to mathematics, but is apparently present in PISA waves when reading was the main assessed cognitive domain. This fact is expressed in R2 values reported in Table 6.2.6, and so it is graphically presented in Figures 6.2.28-6.2.32.

Table 6.2.6 Countries with high levels of negative LID in reading

	R2 FOR Y=1/(A+ BX)	COUNTRY	Value of Possible Outlier	ESD |Z|	Grubbs' Single-Outlier Level Test Prob Level (Alternative Hypothesis: One-Sided vs Maximum)	Conclude Outlier by Rosner's Procedure	Tukey, 1977 - Outside values
PISA 2000	0.8	**Japan (JPN)**	3.7	2.60	**0.059**	**Yes**	**Yes**
		Greece (GRE)	2.3	2.05	0.377	No	No
PISA 2003	0.3	**Greece (GRE)**	9.4	2.50	**0.089**	**Yes**	No
		Austria (AUT)	6.1	2.01	0.426	No	No
PISA 2006	0.43	**Denmark (DNK)**	7.8	2.34	0.157	No	**Yes**
		Poland (POL)	6.3	2.27	0.188	No	No
PISA 2009	0.77	**Greece (GRE)**	6.0	2.80	**0.025**	**Yes**	**Yes**
		Italy (ITA)	4.2	2.51	**0.079**	**Yes**	No
PISA 2012	0.24	**Greece (GRE)**	14.6	3.14	**0.004**	**Yes**	**Yes**
		Italy (ITA)	7.0	2.16	0.273	No	No

Grey areas highlight the PISA waves for which reading was a targeted domain.

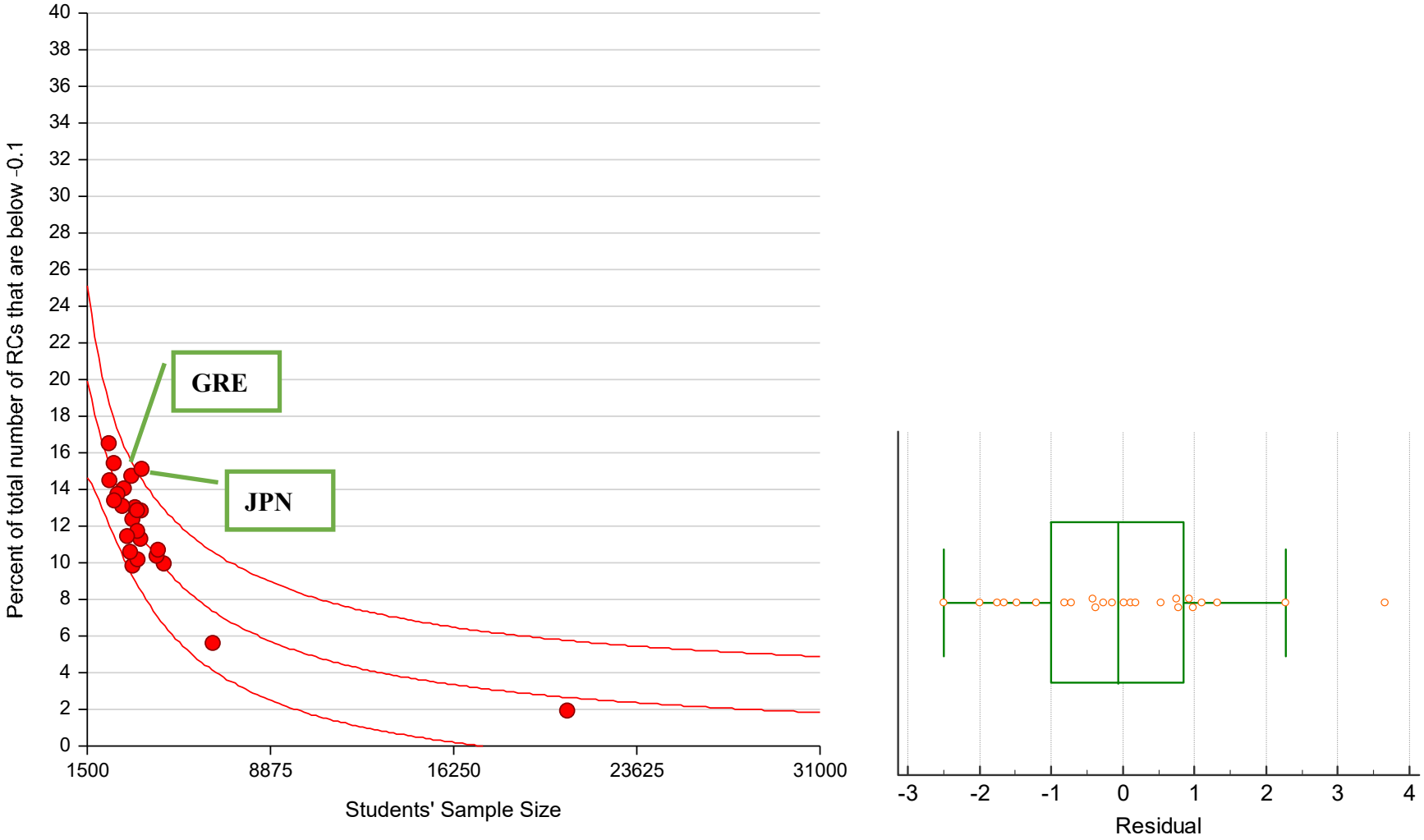


Figure 6.2.28 Reciprocal function and its prediction limits fitted to picture association between student sample size and prevalence of RCs along with a box-plot of prediction residuals (PISA 2000 / RCs that are below -0.1 / Reading)

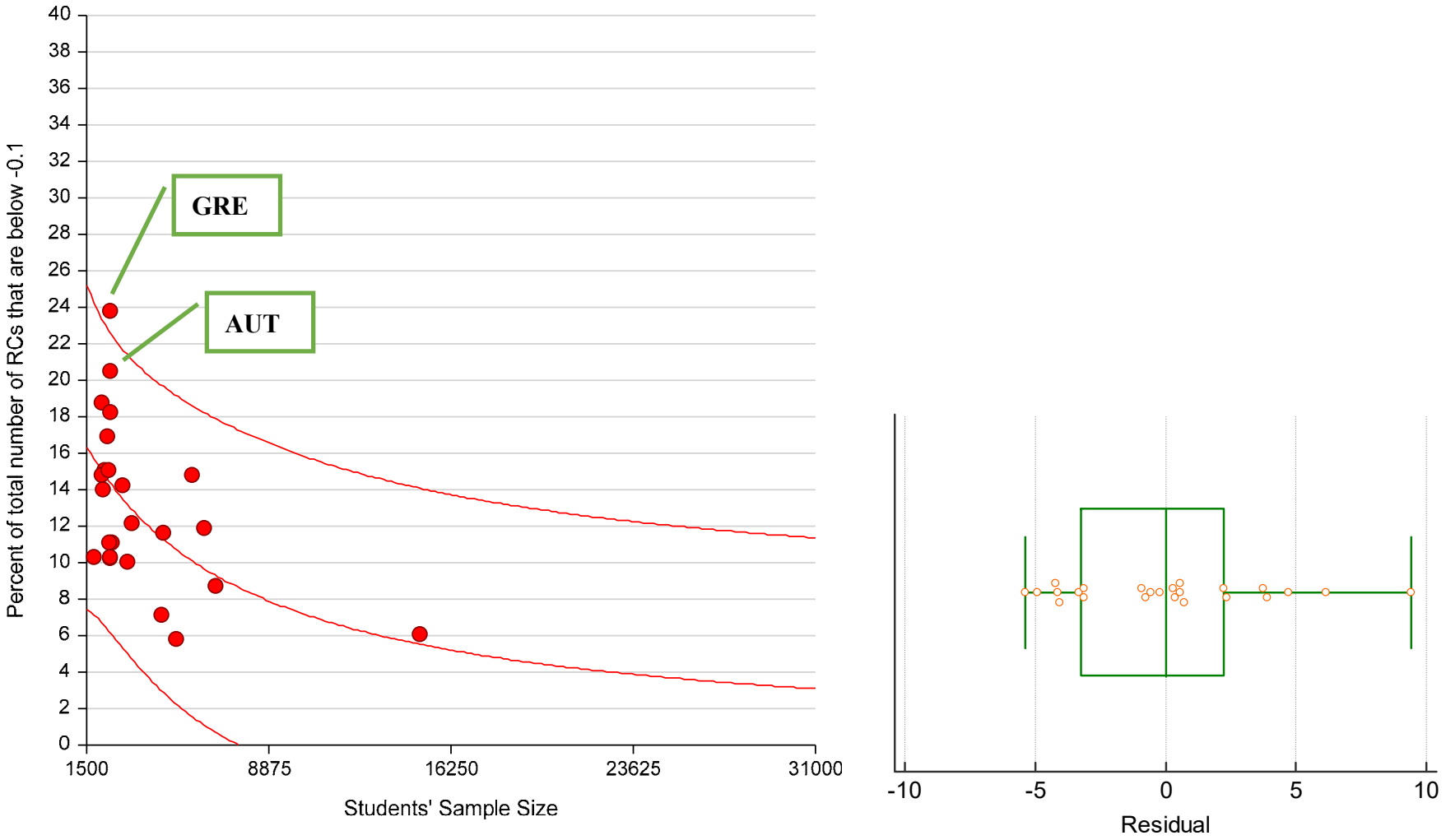


Figure 6.2.29 Reciprocal function and its prediction limits fitted to picture association between student sample size and prevalence of RCs along with a box-plot of prediction residuals (PISA 2003 / RCs that are below -0.1 / Reading)

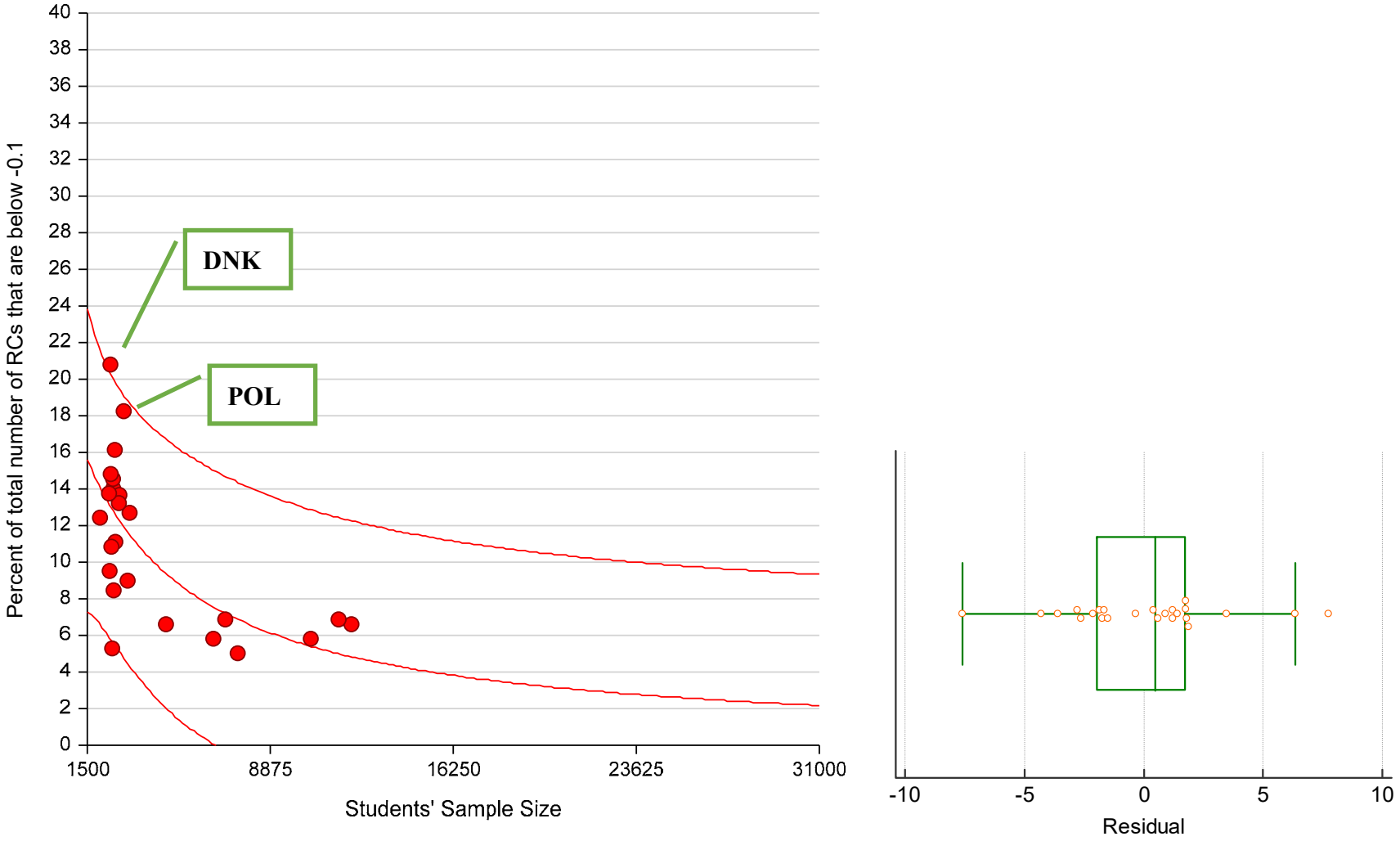


Figure 6.2.30 Reciprocal function and its prediction limits fitted to picture association between student sample size and prevalence of RCs along with a box-plot of prediction residuals (PISA 2006 / RCs that are below -0.1 / Reading)

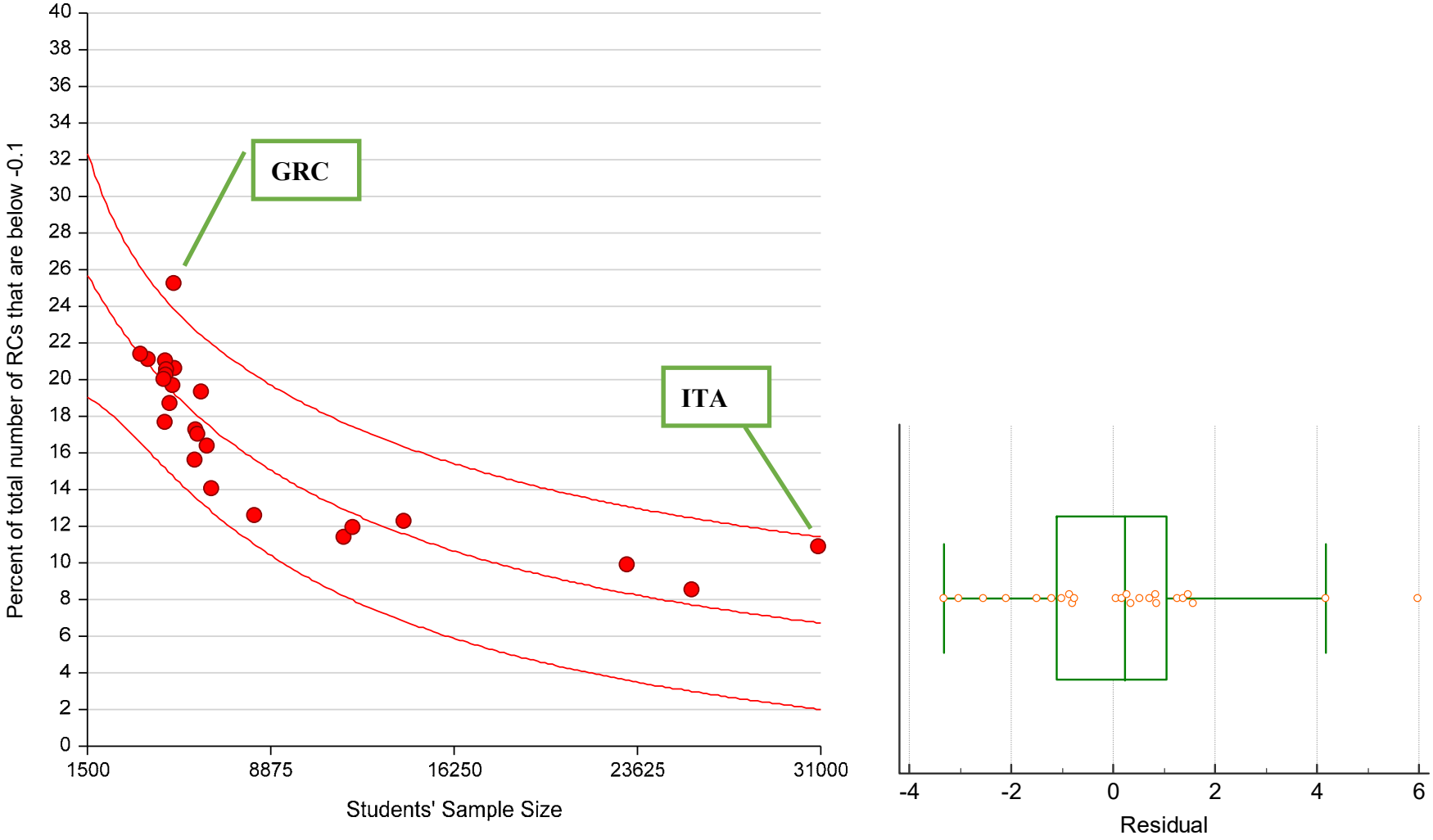


Figure 6.2.31 Reciprocal function and its prediction limits fitted to picture association between student sample size and prevalence of RCs along with a box-plot of prediction residuals (PISA 2009 / RCs that are below -0.1 / Reading)

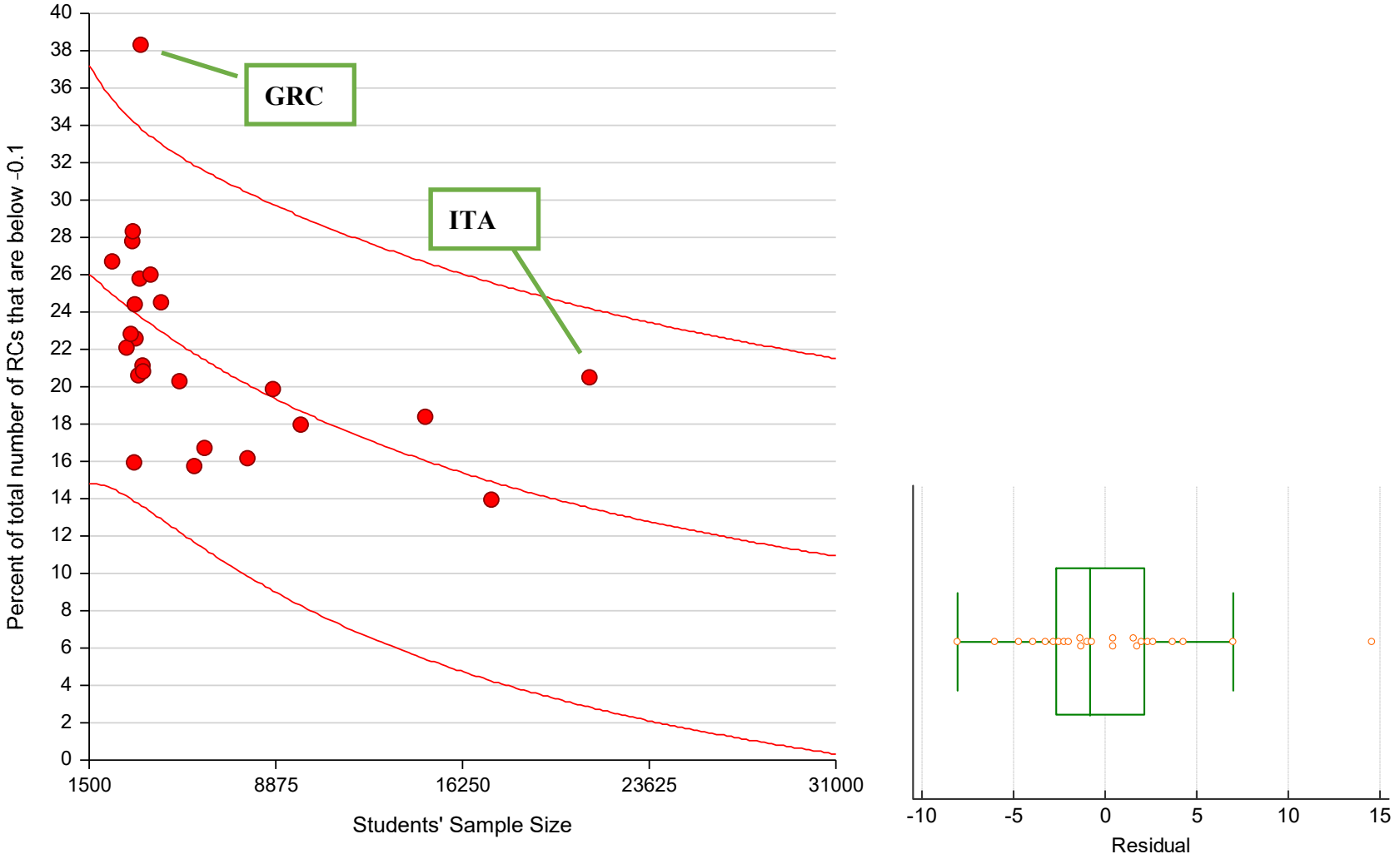


Figure 6.2.32 Reciprocal function and its prediction limits fitted to picture association between student sample size and prevalence of RCs along with a box-plot of prediction residuals (PISA 2012 / RCs that are below -0.1 / Reading)

As can be seen in Table 6.2.6, in PISA 2000, Japan (JPN) revealed a high percent of negative LID while in PISA 2003, 2009 and 2012 Greece featured once again as an outlier among the 24 OECD countries. For reading, Italy (ITA) also revealed high proportions of residual correlations below -0.1.

6.2.5 National calibrations with positive LID in science

In line with previous sections describing mathematics and reading, the discussion about national LID prevalence results for science follows the same template. Figure 6.2.33 presents a dual graph showing the percentage of science item pair RCs exceeding 0.1 out of all item pair RCs against students' sample sizes. Once again a negative relationship between the percent of RCs above the cut-point and the size of students' cohorts is present and is most clearly evident in PISA 2006. In this year, when science was the targeted cognitive domain, Finland's (FIN) 11% percent of the pairs of items indicating positive LID appeared to be higher when compared to the larger sampled Japan (JPN) [48] and Czech Republic (CZE) [49],both reporting at approximately 7.2%. As it can be inferred from Figure 6.2.34, this discrepancy is due to pairs of items from different testlets. It is also of interest to highlight the consistently small presence of within-testlet positive LID despite the majority of testlets used in PISA's science evaluation incorporating three and four items in each testlet, as can be confirmed in Figure 4.3.3.

[48] Non-corrected for multiple comparisons 95%CIs for the proportion difference is (2.7%-4.9%). Epitools were used - http://epitools.ausvet.com.au/content.php?page=z-test-2&p1=0.072&p2=0.11&n1=5151&n2=5253&Conf=0.05&tails=2&samples=2

[49] Non-corrected for multiple comparisons 95%CIs for the proportion difference is (2.7%-4.9%). Epitools were used - http://epitools.ausvet.com.au/content.php?page=z-test-2&p1=0.072&p2=0.11&n1=5253&n2=5253&Conf=0.05&tails=2&samples=2

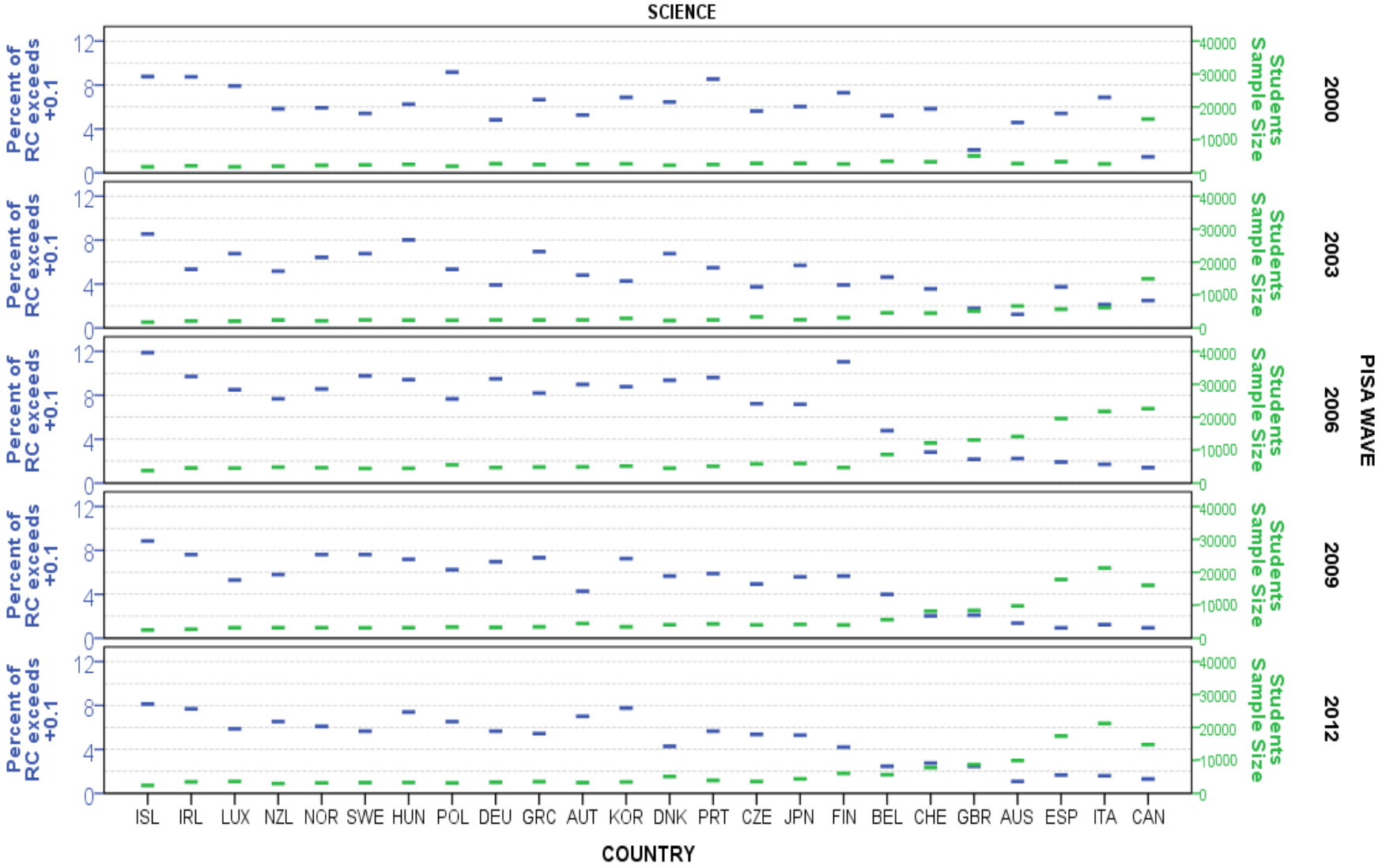


Figure 6.2.33 Dual graph showing the percent of science item pairs with RCs exceeding 0.1 taken out of all RCs against students' sample sizes

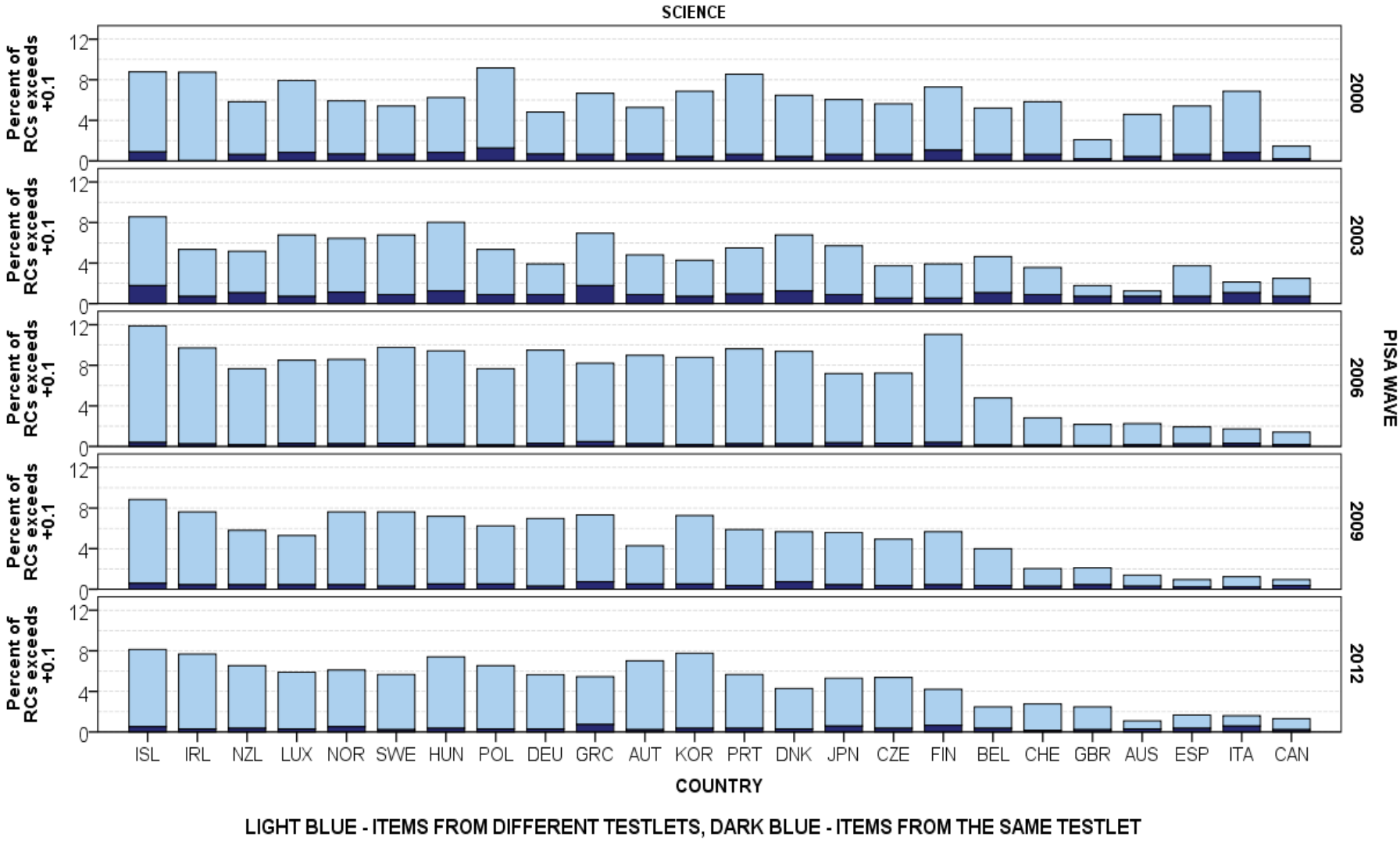


Figure 6.2.34 Percent of science item pairs with RCs exceeding 0.1 taken out of all RCs separated into components involving item pairs from the same testlets and from different testlets

On examining countries with a high percent of between-testlets RCs suggesting positive LID (see Table 6.2.7 and corresponding Figures 6.2.35-6.2.39) in PISA 2006, Finland (FIN) is clearly an outlier, and so are Ireland (IRL) and Korea (KOR) in PISA 2012.

Table 6.2.7 Countries with high levels of between-testlets positive LID in science

	R2 FOR Y=1/(A+BX)	COUNTRY	Value of Possible Outlier	ESD |Z|	Grubbs' Single-Outlier Level Test Prob Level (Alternative Hypothesis: One-Sided vs Maximum)	Conclude Outlier by Rosner's Procedure	Tukey, 1977 - Outside values
PISA 2000	0.66	Portugal (PRT)	2.1	2.08	0.364	No	No
		Ireland (IRL)	2.0	2.29	0.175	No	No
PISA 2003	0.75	Hungary (HUN)	1.9	2.16	0.282	No	No
		Belgium (BEL)	1.2	1.62	1.000	No	No
PISA 2006	0.96	**Finland (FIN)**	1.8	2.76	**0.030**	**Yes**	**Yes**
		Portugal (PRT)	1.3	2.61	**0.053**	**Yes**	No
PISA 2009	0.92	Portugal (PRT)	1.0	1.52	1.000	No	No
		Korea (KOR)	0.9	1.52	1.000	No	No
PISA 2012	0.88	**Ireland (IRL)**	1.7	2.26	**0.208**	**Yes**	**Yes**
		Korea (KOR)	1.7	2.60	**0.055**	**Yes**	**Yes**

Grey areas highlight the PISA waves for which science was a targeted domain.

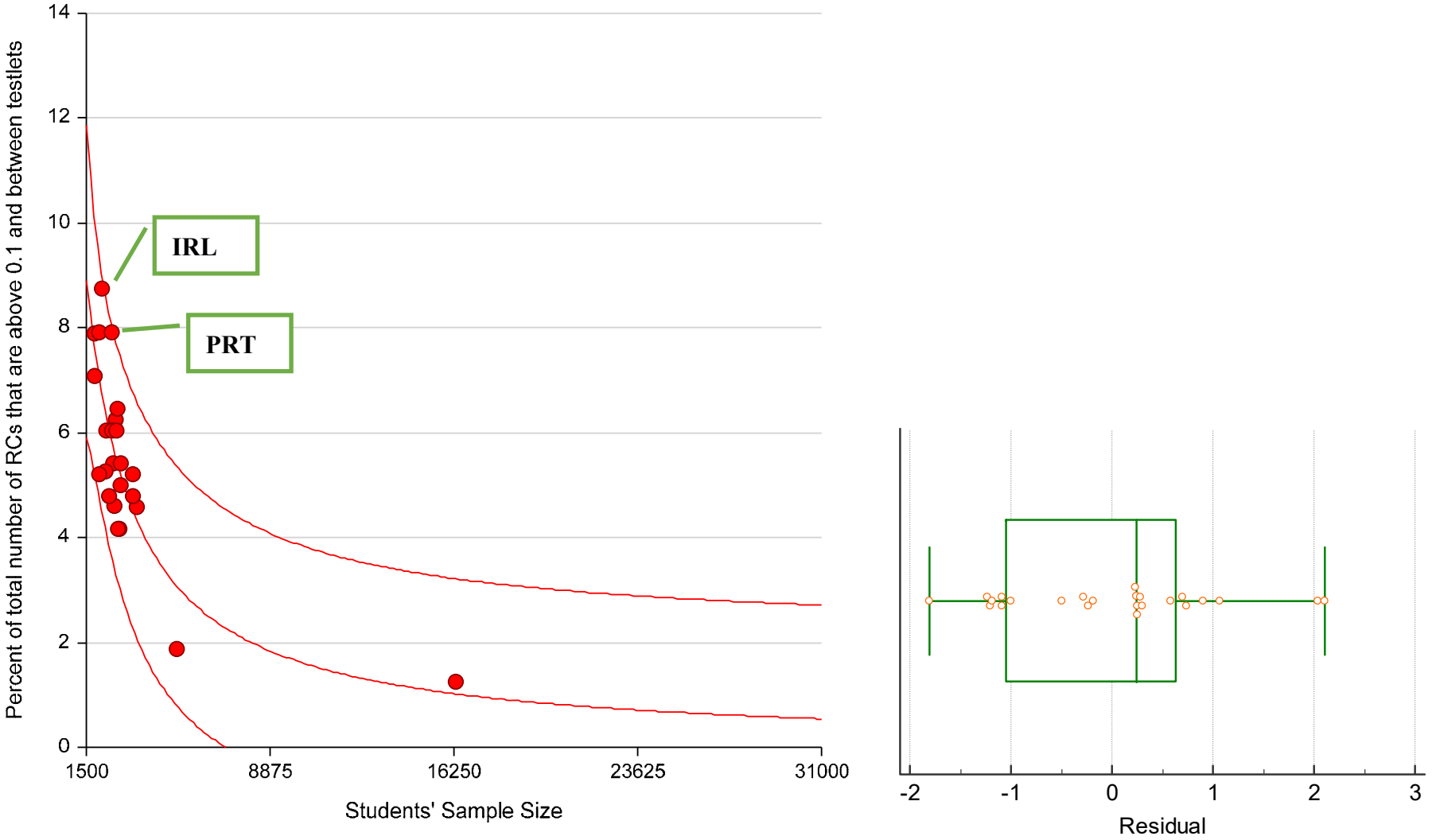


Figure 6.2.35 Reciprocal function and its prediction limits fitted to picture association between student sample size and prevalence of RCs along with a box-plot of prediction residuals (PISA 2000 / RCs that are above 0.1 / Pairs of items from different testlets / Science)

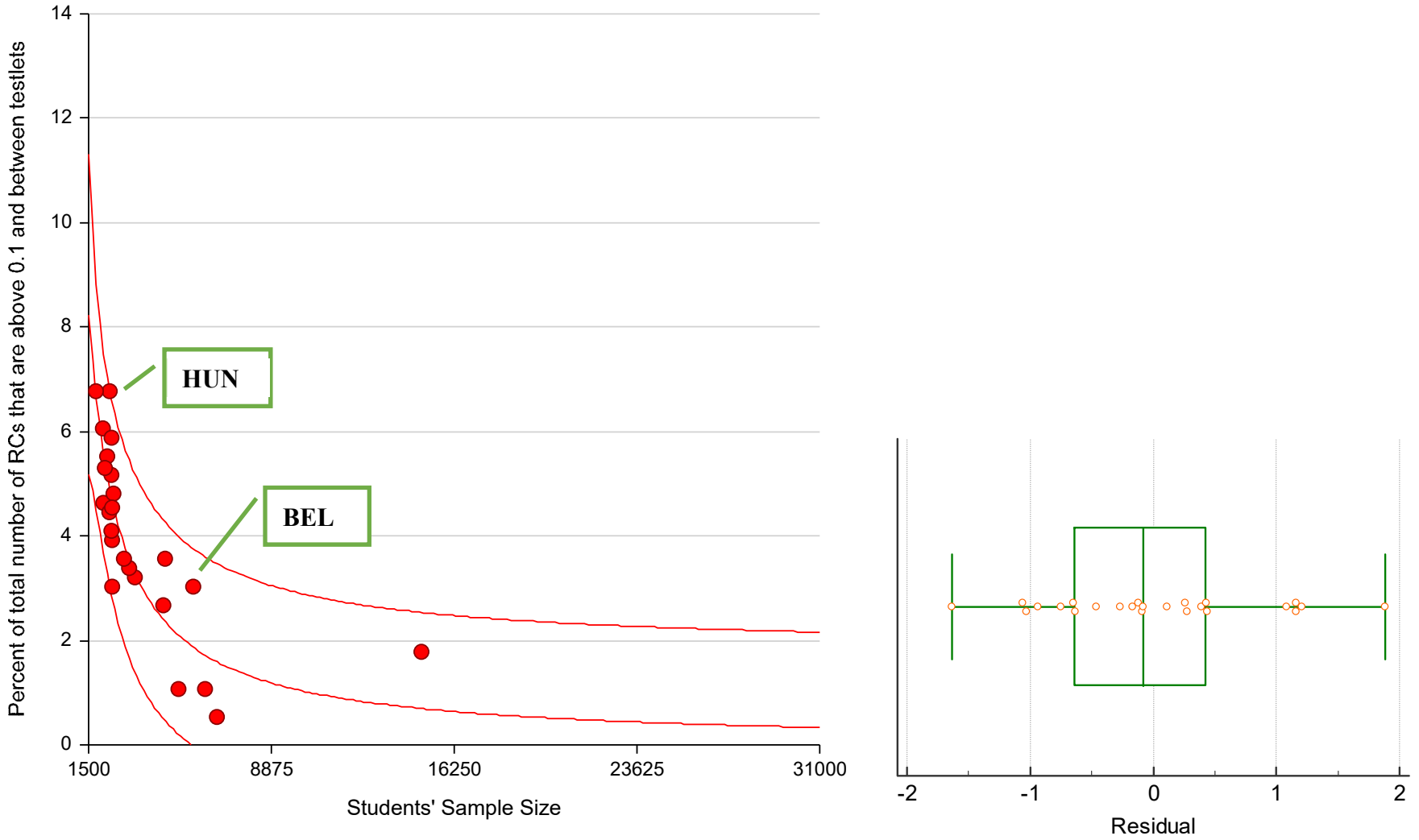


Figure 6.2.36 Reciprocal function and its prediction limits fitted to picture association between student sample size and prevalence of RCs along with a box-plot of prediction residuals (PISA 2003 / RCs that are above 0.1 / Pairs of items from different testlets / Science)

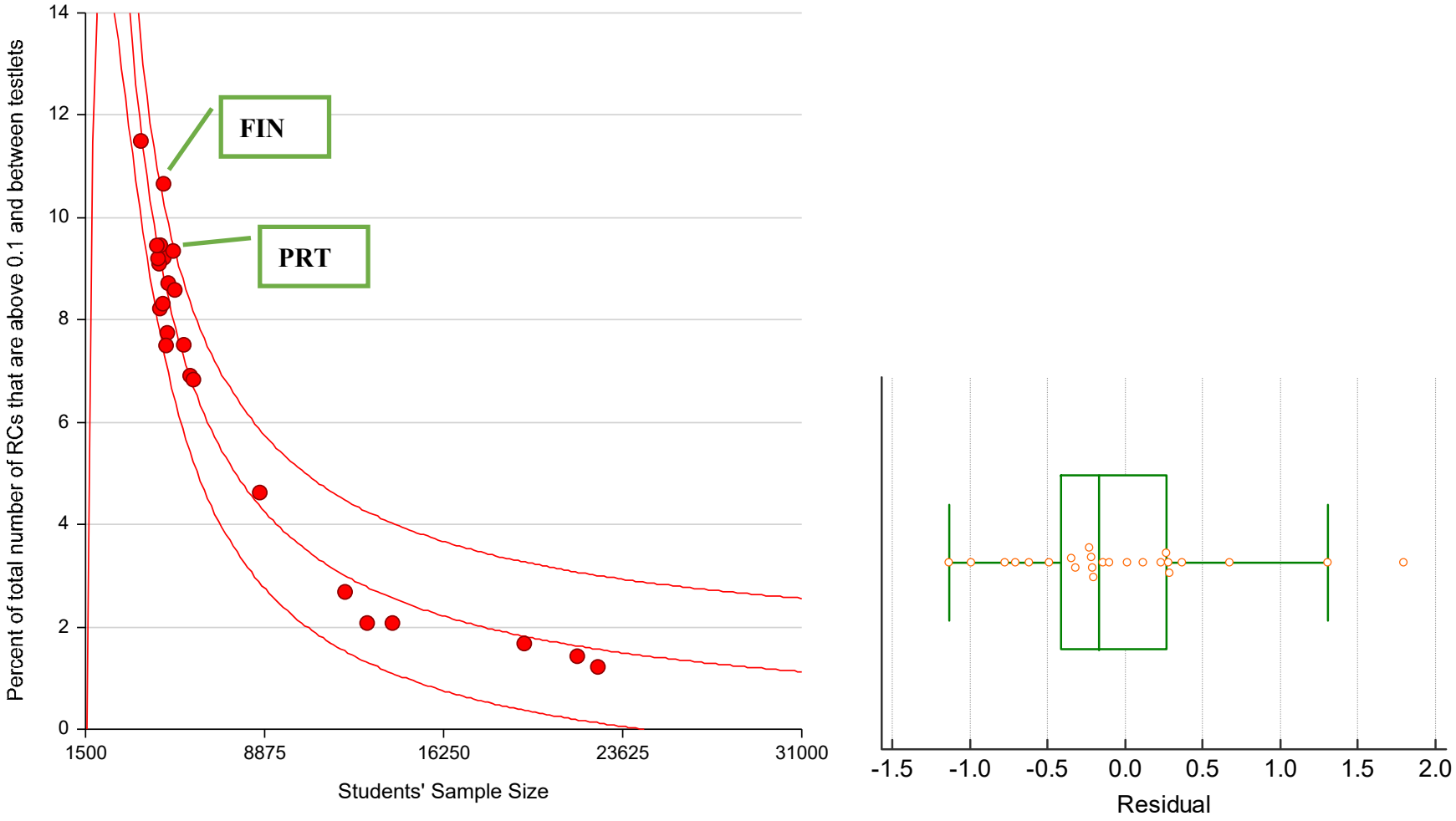


Figure 6.2.37 Reciprocal function and its prediction limits fitted to picture association between student sample size and prevalence of RCs along with a box-plot of prediction residuals (PISA 2006 / RCs that are above 0.1 / Pairs of items from different testlets / Science)

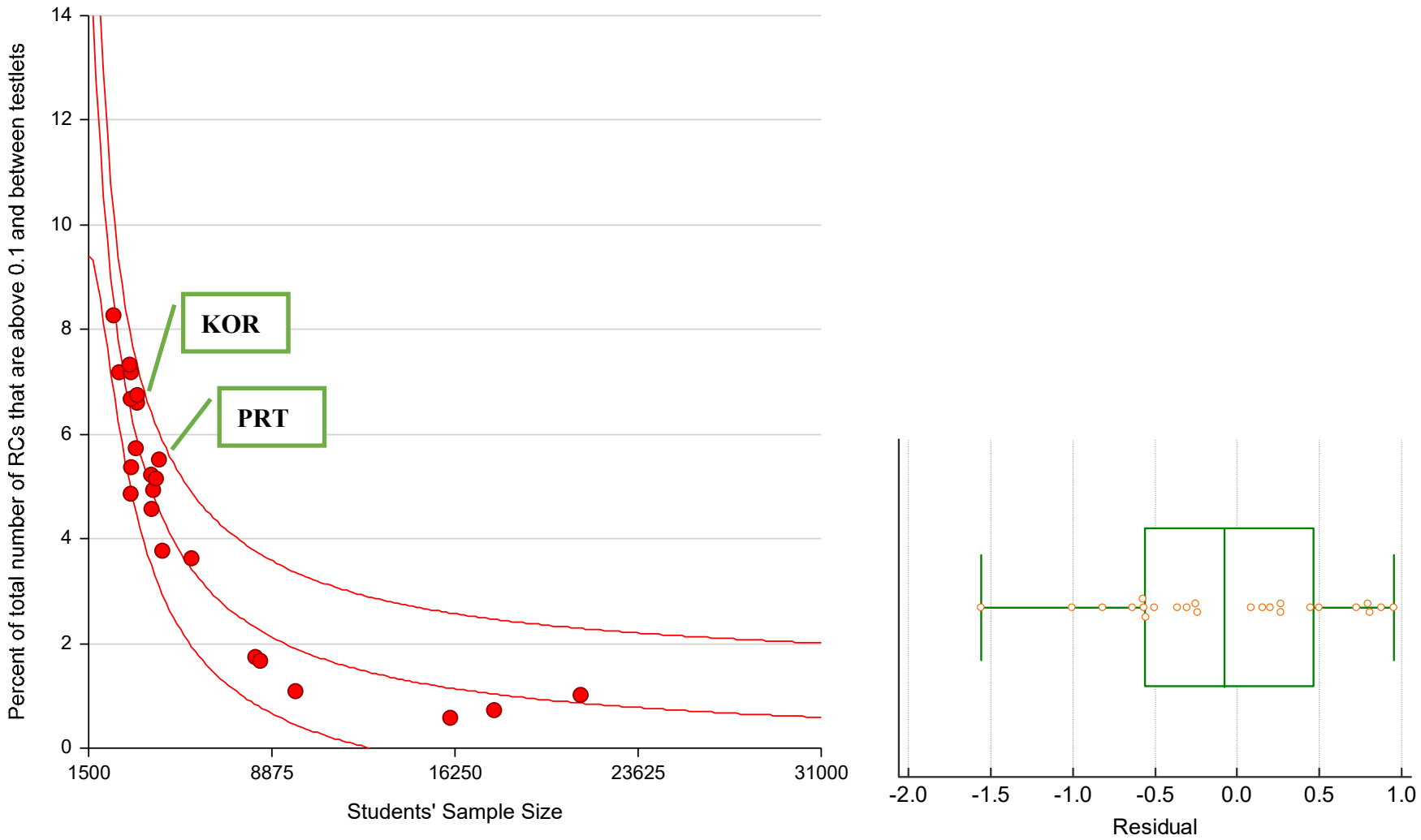


Figure 6.2.38 Reciprocal function and its prediction limits fitted to picture association between student sample size and prevalence of RCs along with a box-plot of prediction residuals (PISA 2009 / RCs that are above 0.1 / Pairs of items from different testlets / Science)

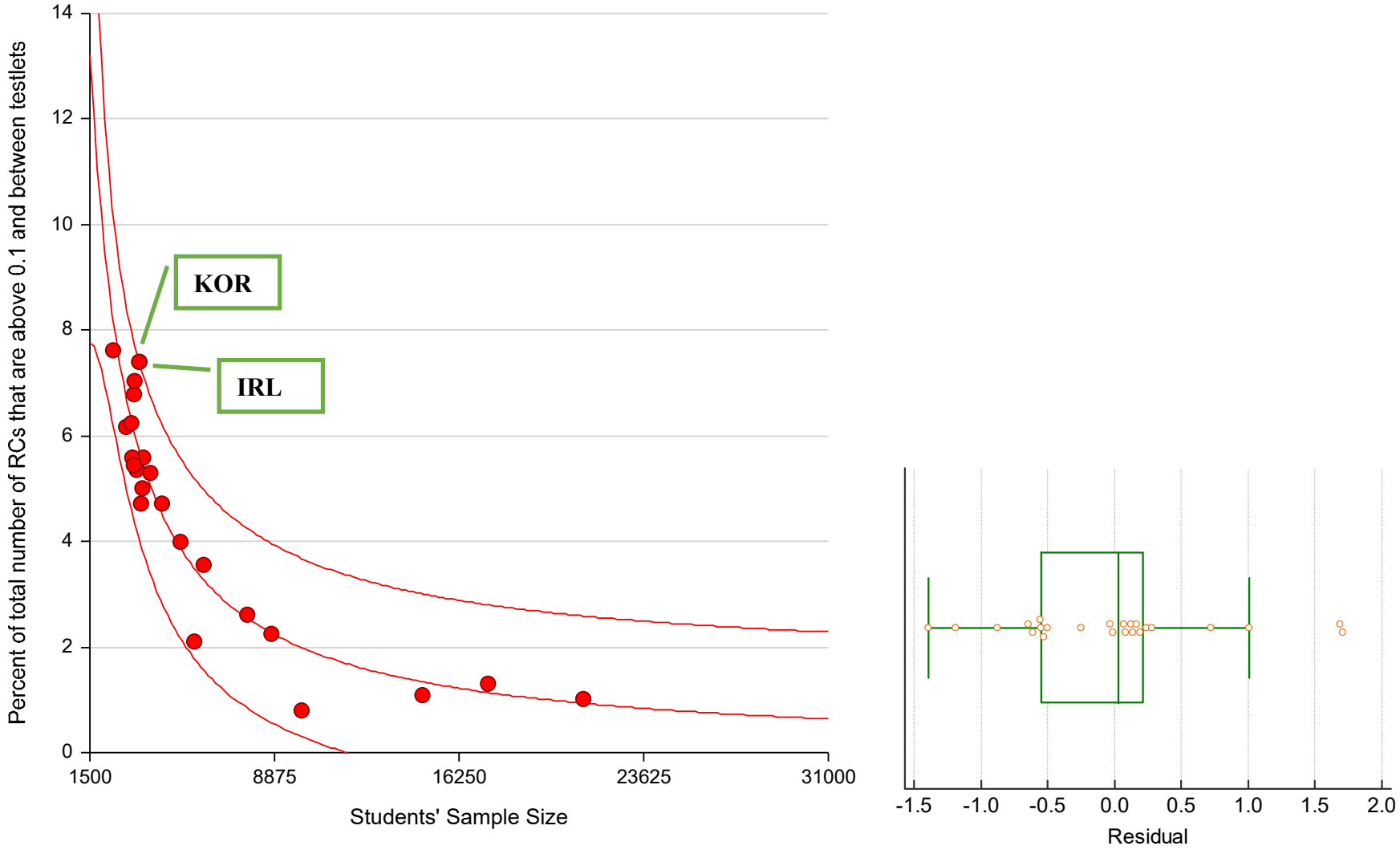
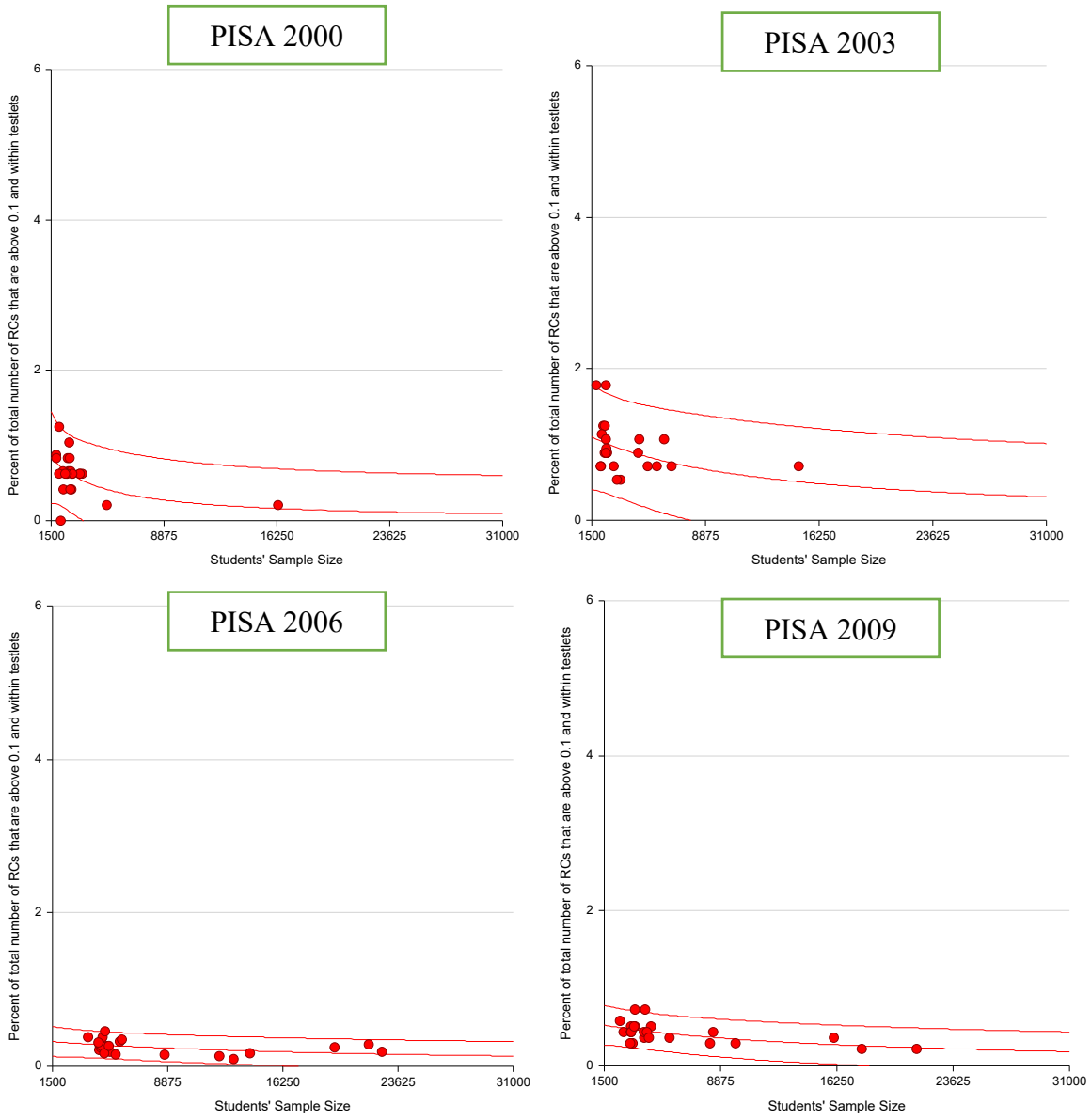


Figure 6.2.39 Reciprocal function and its prediction limits fitted to picture association between student sample size and prevalence of RCs along with a box-plot of prediction residuals (PISA 2012 / RCs that are above 0.1 / Pairs of items from different testlets / Science)

Set of countries with a high proportion of within-testlet RCs exceeding 0.1 was observed. This can be seen in Figure 6.2.40 and concluding this figure Table 6.2.8 represents the results of identifying the outliers of prevalence estimates.



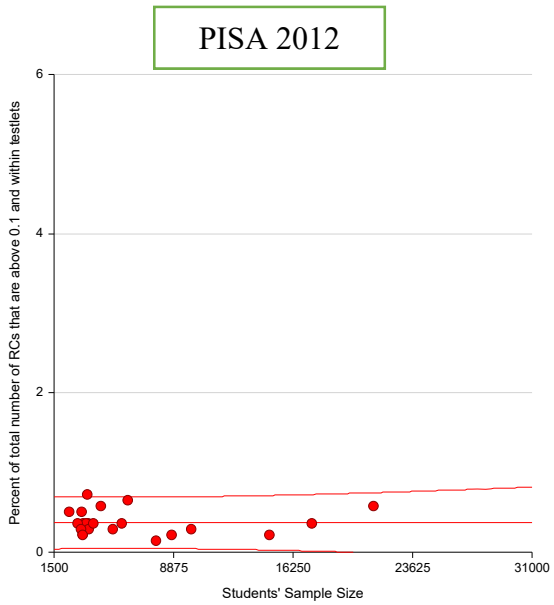


Figure 6.2.40 Reciprocal function and its prediction limits fitted to show the association between students' sample size and prevalence of (PISA 2000,2003,2006,2009 and 2012 / RCs that are above 0.1 / Pairs of items from within the same testlets / Science)

Table 6.2.8 Countries with high levels of within-testlet positive LID in science

	COUNTRY	Value of Possible Outlier	ESD|Z|	Grubbs' Single-Outlier Level Test Prob Level (Alternative Hypothesis: One-Sided vs Maximum)	Conclude Outlier by Rosner's Procedure	Tukey, 1977 - Outside values
PISA 2000	**Poland (POL)**	1.3	2.36	0.148	No	**Yes**
	Finland (FIN)	1.0	1.90	0.571	No	No
PISA 2003	**Greece (GRC)**	1.8	2.58	**0.064**	**Yes**	**Yes**
	Iceland (ISL)	1.8	3.16	**0.003**	**Yes**	**Yes**
PISA 2006	Greece (GRC)	0.5	2.32	0.171	No	No
	Finland (FIN)	0.4	1.74	0.848	No	No
	Iceland (ISL)	0.4	1.92	0.504	No	No
PISA 2009	**Denmark (DNK)**	0.7	2.27	**0.199**	**Yes**	No
	Greece (GRC)	0.7	2.65	**0.045**	**Yes**	No
PISA 2012	**Greece (GRC)**	0.7	2.39	0.134	No	**Yes**
	Finland (FIN)	0.7	2.28	0.184	No	**Yes**

Grey areas highlight the PISA waves for which science was a targeted domain.

Once again Greece (GRC) featured in Table 6.2.8 in multiple PISA waves, although only in PISA 2003 was it identified as an outlier according to all detection methods. Iceland (ISL) also had higher levels of within-testlet positive LID in PISA 2003. Mimicking previous discussions about mathematics and reading, Figure 6.2.41 reports the percentages of within-testlet RC exceeding 0.1 when the denominator includes only pairs of items within the testlets.

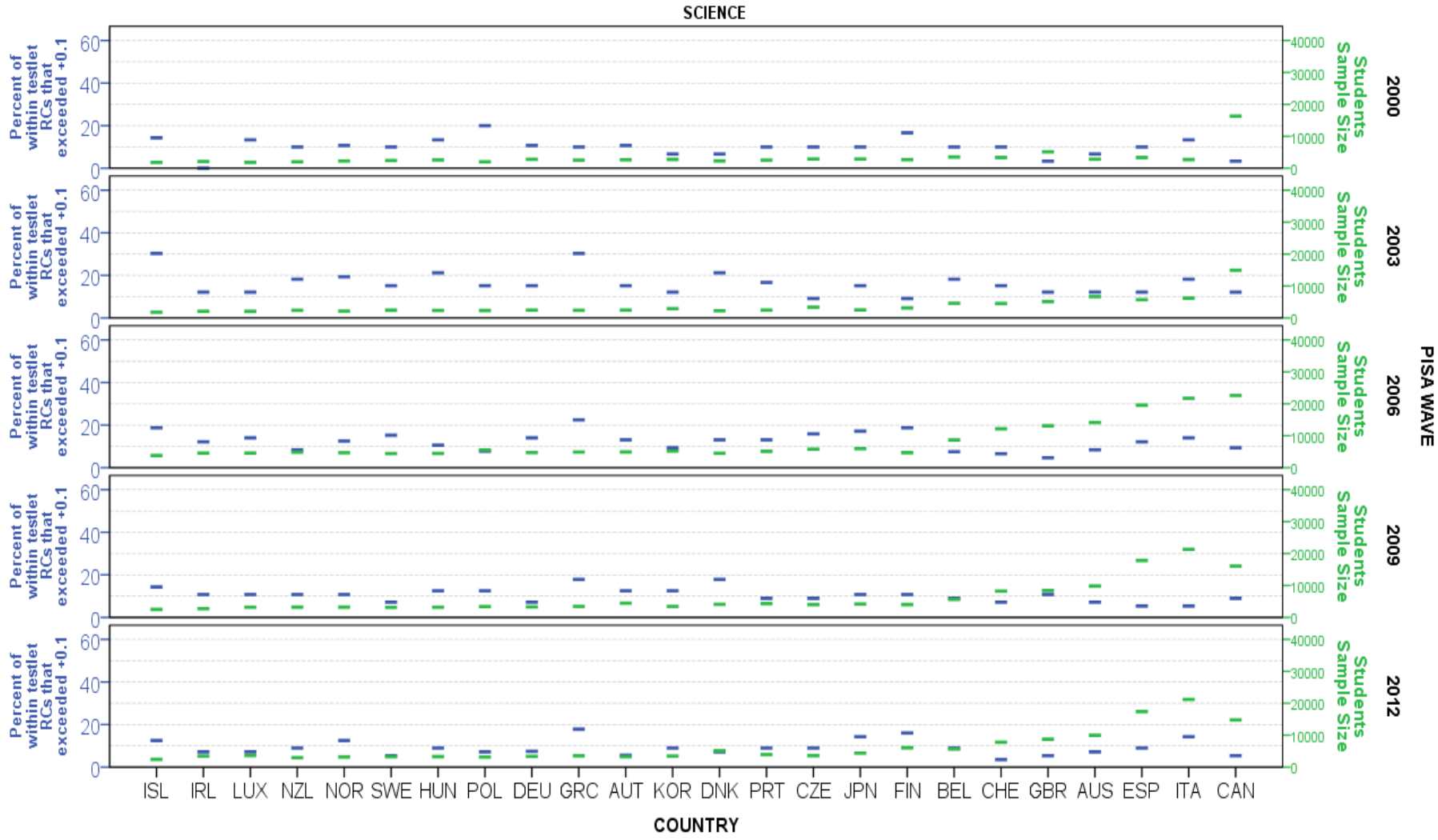


Figure 6.2.41 Dual graph showing the percent of science item pairs with RCs exceeding 0.1 taken out of total of only within-testlet RCs against students sample sizes

The prevalence of within-testlet positive LID seems to be smaller, as compared to previously discussed cognitive domains. This can be seen in Figure 6.2.42. The science domain became a targeted literacy only in PISA 2006.

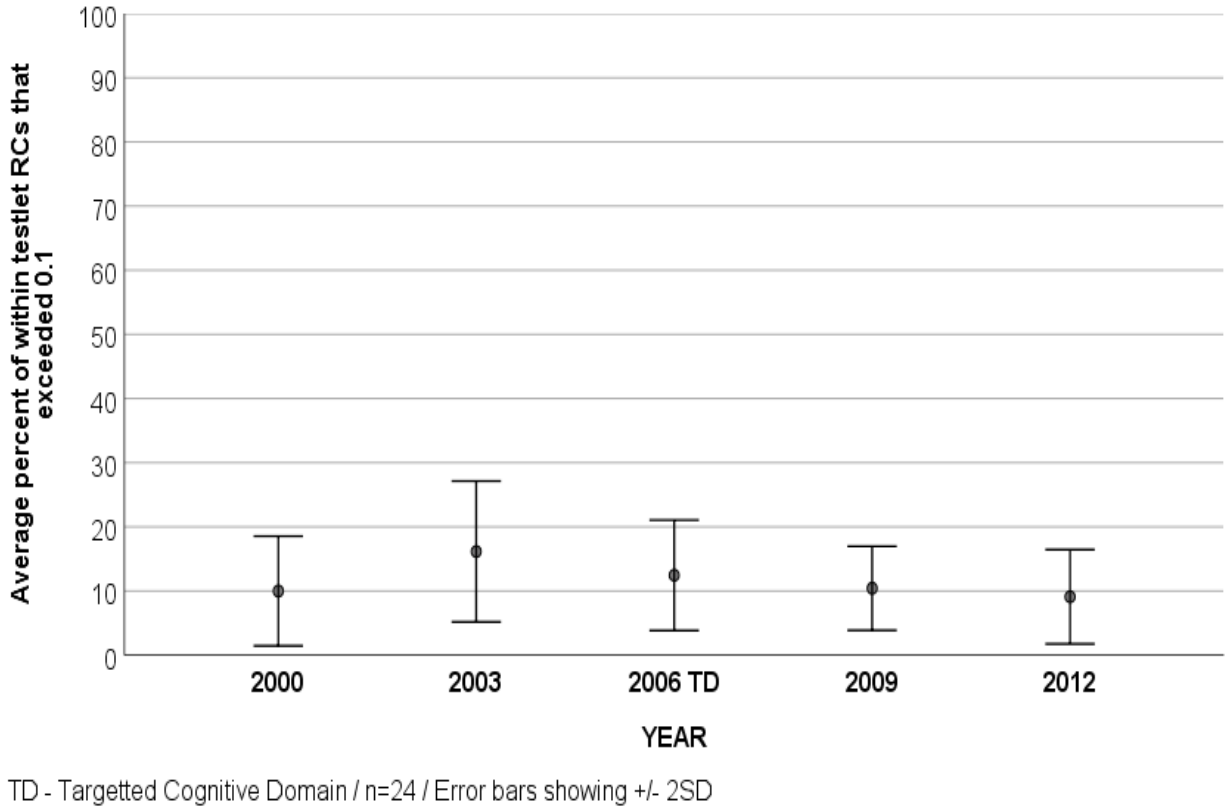


Figure 6.2.42 Average percentage of science item pairs with RCs exceeding 0.1 of total within-testlet RCs obtained from 24 OCED countries.

6.2.6 National calibrations with negative LID in science

As can be seen in Figure 6.2.43, particularly in the science targeted PISA 2006, the relationship between the prevalence of below -0.1 RCs is similar to the previously observed corresponding figures in mathematics and reading.

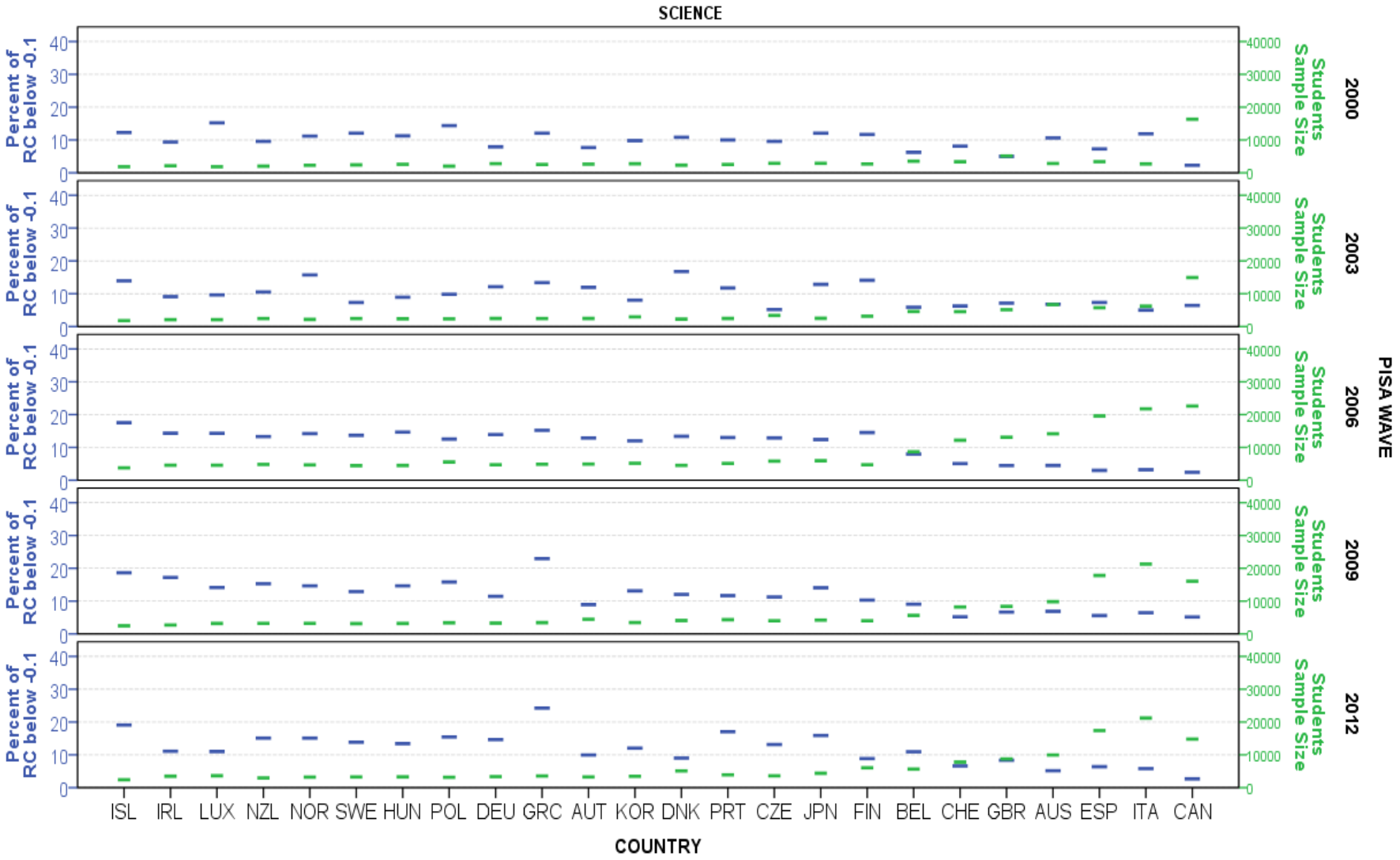


Figure 6.2.43 Dual graph showing the percent of science item pairs with RCs below -0.1 taken out of all RCs against students' sample sizes

Table 6.2.9 and Figures 6.2.44-6.2.48 suggest that Greece (GRC) has higher proportions of RCs below -0.1 compared to other countries.

Table 6.2.9 Countries with high levels of negative LID in science

	R2 FOR Y=1/(A +BX)	COUNTRY	Value of Possible Outlier	ESD |Z|	Grubbs' Single-Outlier Level Test Prob Level (Alternative Hypothesis: One-Sided vs Maximum)	Conclude Outlier by Rosner's Procedure	Tukey, 1977 - Outside values
PISA 2000	0.7	Japan (JPN)	2.7	1.68	1.000	No	No
		Italy (ITA)	2.0	1.37	1.000	No	No
PISA 2003	0.44	Denmark (DNK)	5.1	1.96	0.511	No	No
		Finland (FIN)	4.4	1.92	0.532	No	No
PISA 2006	0.98	**Greece (GRC)**	1.6	2.30	0.178	No	**Yes**
		Czech Rep (CZE)	1.4	2.40	0.120	No	**Yes**
PISA 2009	0.71	**Greece (GRC)**	8.8	3.51	**0.000**	**Yes**	**Yes**
		Italy (ITA)	3.5	2.22	0.221	No	No
PISA 2012	0.62	**Greece (GRC)**	10.2	3.35	**0.001**	**Yes**	**Yes**
		Portugal (PRT)	3.8	1.95	0.491	No	No

Grey areas highlight the PISA waves for which science was a targeted domain.

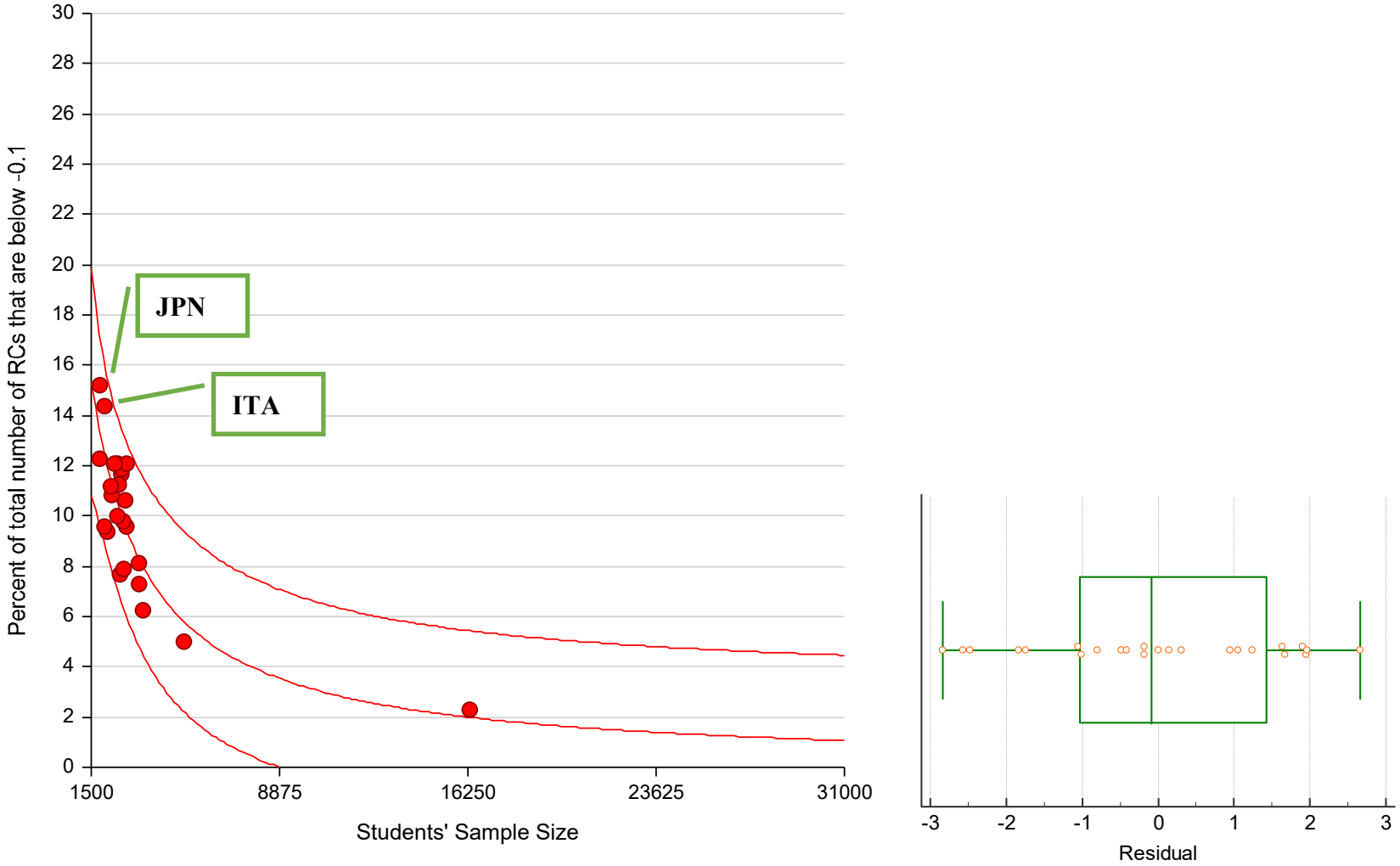


Figure 6.2.44 Reciprocal function and its prediction limits fitted to picture association between student sample size and prevalence of RCs along with a box-plot of prediction residuals (PISA 2000 / RCs that are below -0.1 / Science)

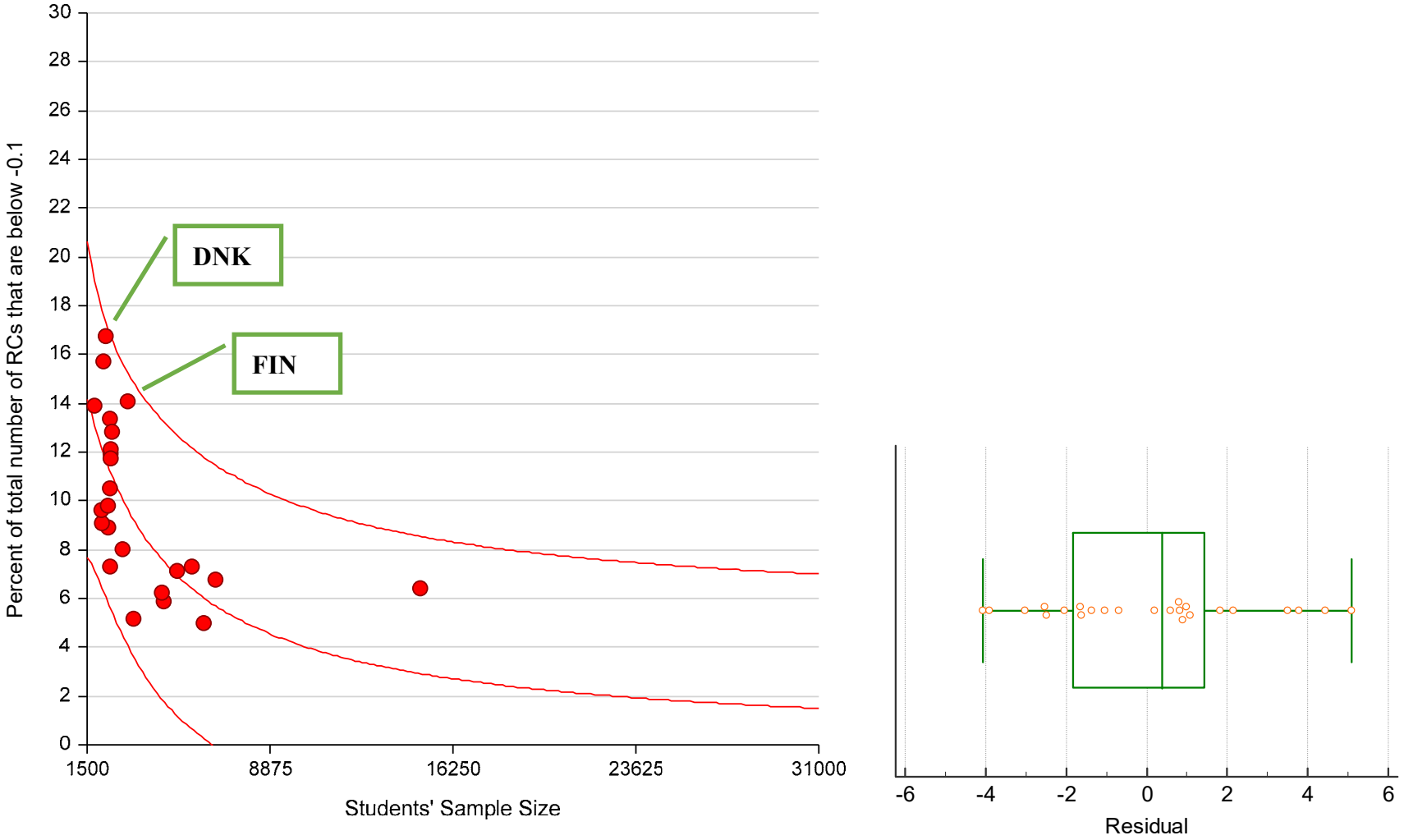


Figure 6.2.45 Reciprocal function and its prediction limits fitted to picture association between student sample size and prevalence of RCs along with a box-plot of prediction residuals (PISA 2003 / RCs that are below -0.1 / Science)

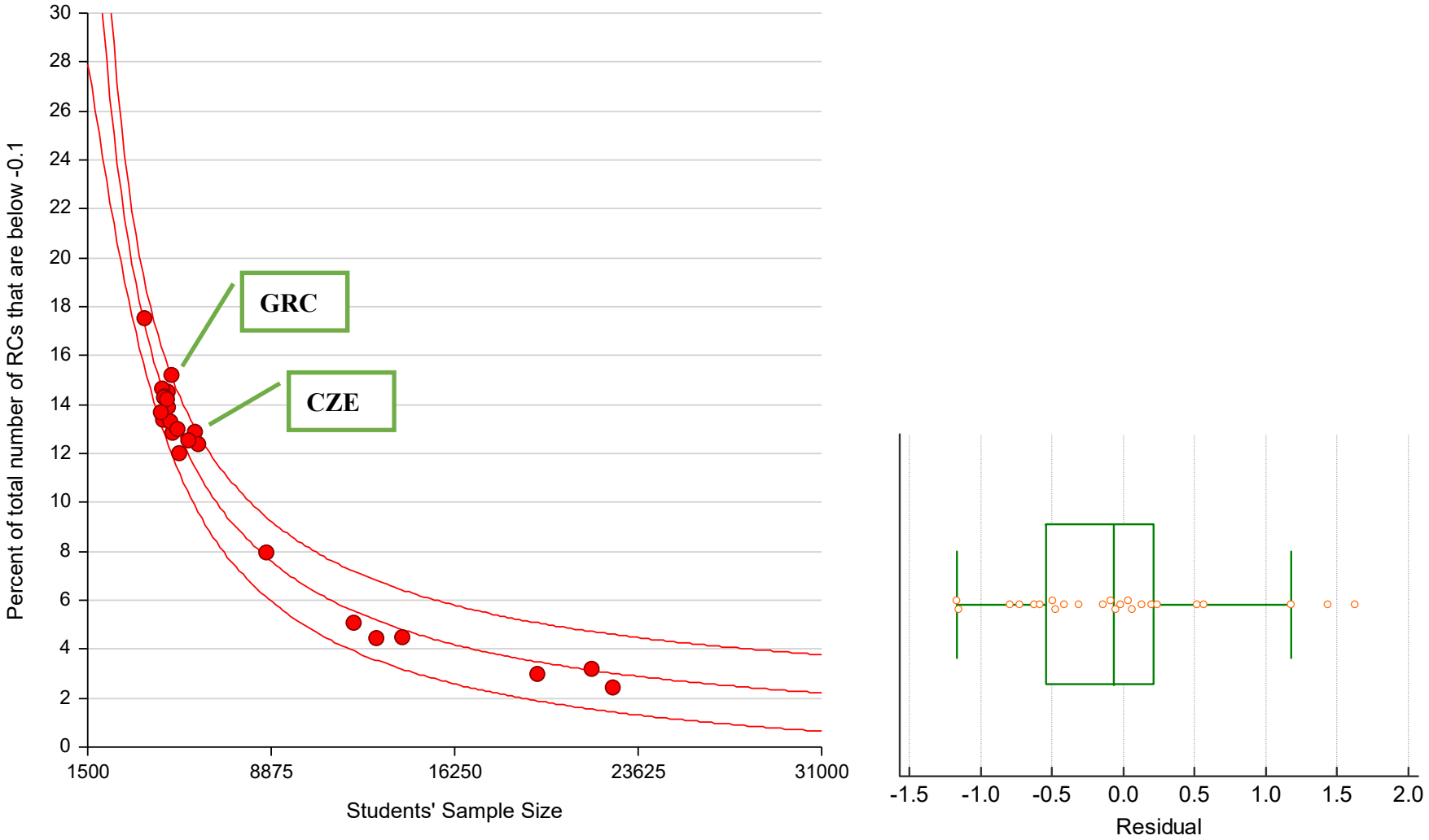


Figure 6.2.46 Reciprocal function and its prediction limits fitted to picture association between student sample size and prevalence of RCs along with a box-plot of prediction residuals (PISA 2006 / RCs that are below -0.1 / Science)

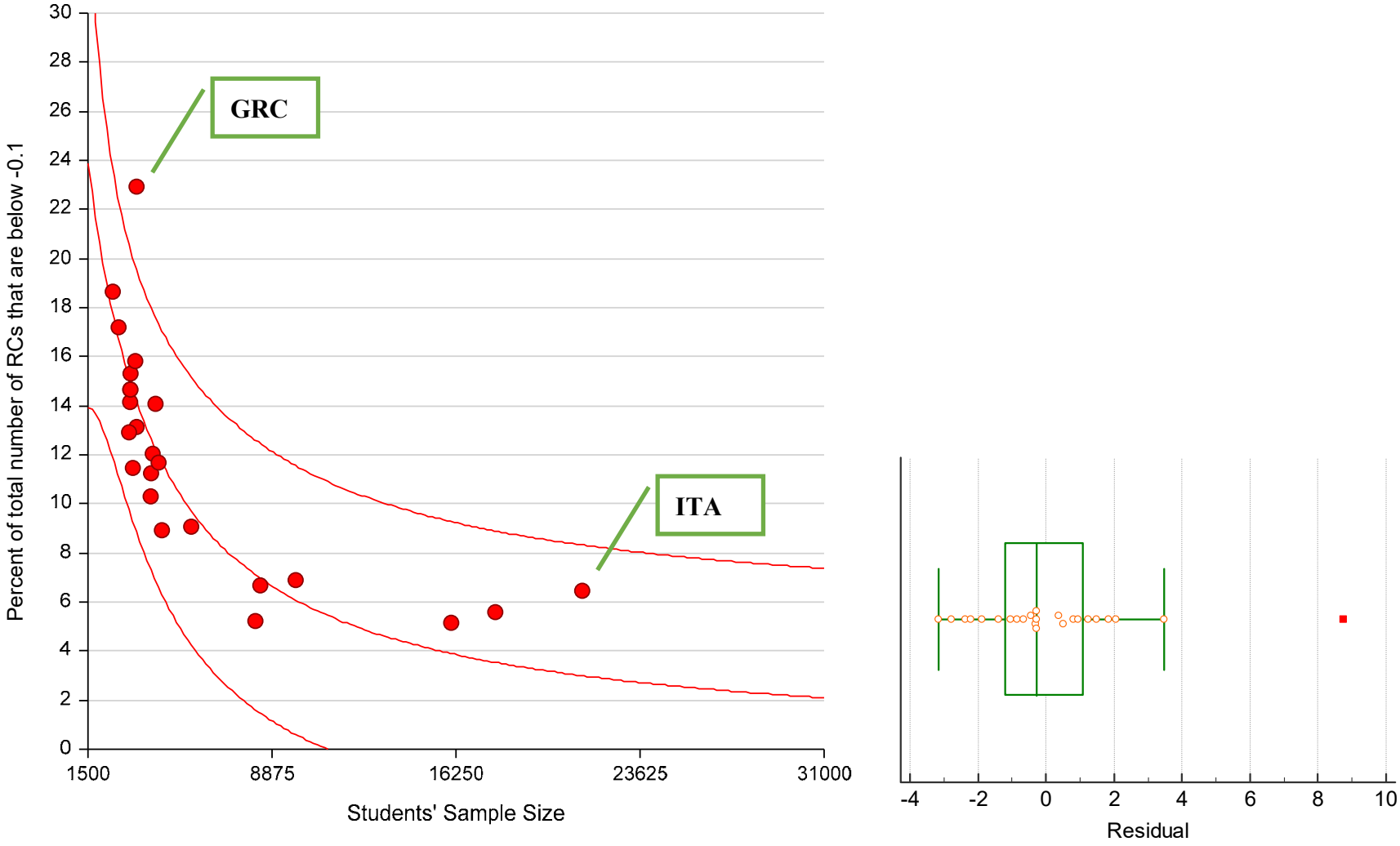


Figure 6.2.47 Reciprocal function and its prediction limits fitted to picture association between student sample size and prevalence of RCs along with a box-plot of prediction residuals (PISA 2009 / RCs that are below -0.1 / Science)

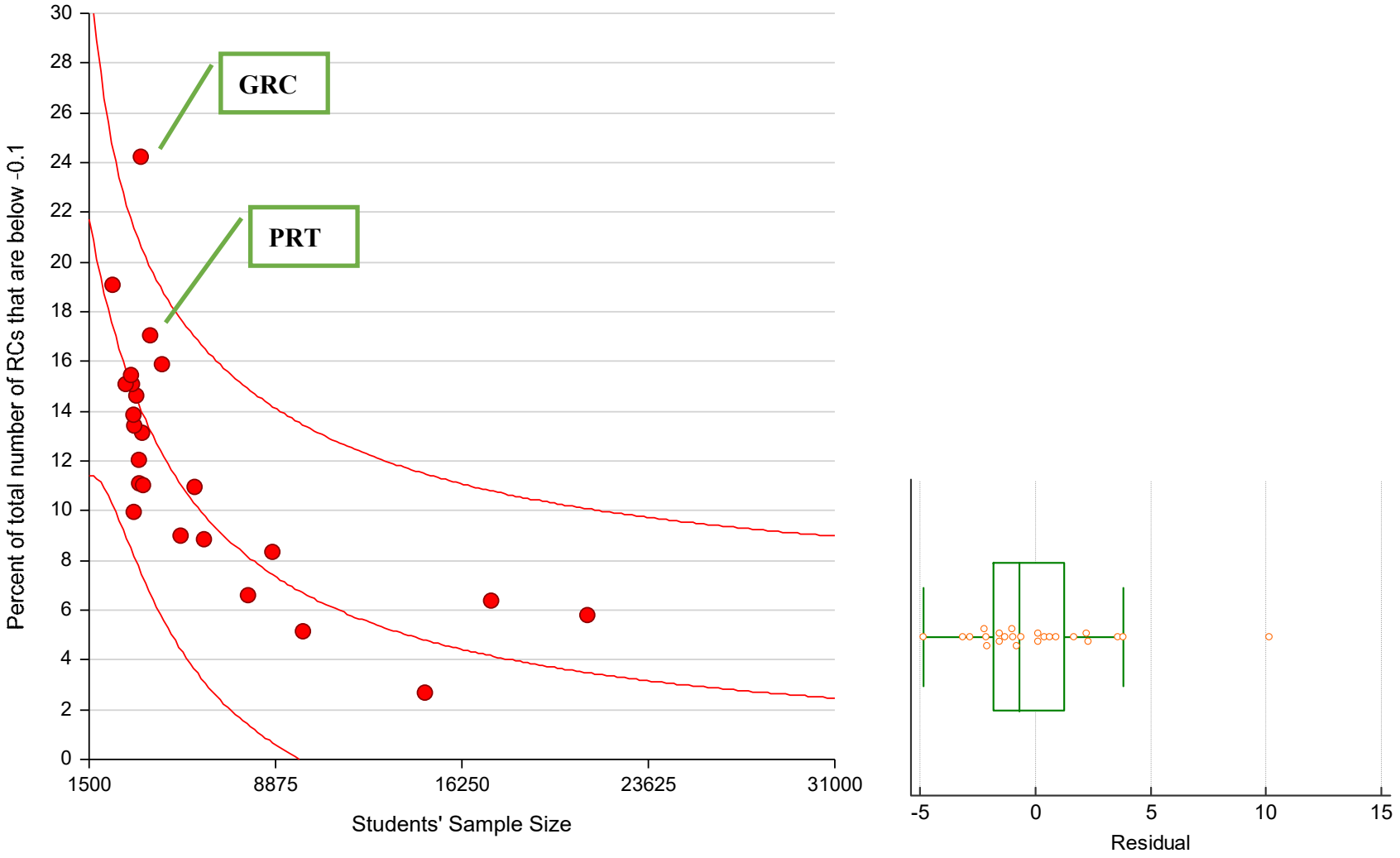


Figure 6.2.48 Reciprocal function and its prediction limits fitted to picture association between student sample size and prevalence of RCs along with a box-plot of prediction residuals (PISA 2012 / RCs that are below -0.1 / Science)

6.2.7 Summary

Addressing the research question 3A and 3B proved to be more challenging than initially anticipated, due to the observed relationship between students' sample sizes and the percentage of RCs either below -0.1 or exceeding 0.1. It is likely that a chosen fixed cut-point of 0.1 will be too liberal for countries with smaller cohorts of students tested in PISA, for example, Iceland (ICE) or Luxembourg (LUX). This cut-point may also be too conservative for indicating LID when very large cohorts of students were used as in the case of Canada (CAN) and Italy (ITA). A paper, recently published by Christensen et al. (2017), offers a suggestion of finding more empirically suitable thresholds by utilising a large number (10000) of simulated datasets. This approach was not possible to utilise in this research given the number of 360 primary CFAs estimations providing the RCs for this part of the project and the estimation time that was required for CFA models that involved targeted cognitive domains. At the same time, using the reciprocal function inferred from Christensen et al. (2017) offered a tool for controlling for the different sizes of students' cohorts. This function fitted the data very well for main targeted domains with the R2 being 0.98 and 0.95 for mathematics (Table 6.2.1), 0.78 and 0.95 for reading (Table 6.2.4), 0.96 for science (Table 6.2.7) when looking at items from different testlets. The function also did well with negative RCs with R2 equal to 0.89 and 0.88 for mathematics (Table 6.2.3), 0.8 and 0.77 for reading (Table 6.2.6) and 0.98 for science (Table 6.2.9). This offers an opportunity to suggest an extension of the Christensen et al. (2017) results to investigate whether a mathematical solution to the problem of a suitable cut-point can be offered, bypassing a need for simulations.

Results presented in Figures 6.2.10, 6.2.26 and 6.2.42 suggest that in mathematics the dependency among items from the same testlets may be more pronounced as compared to reading or science testlets. These figures also suggest the choice of testlets matter with the first two waves of PISA using testlets that show more within-testlet dependency compared to later waves. This result adheres to the international calibration investigation pictorially presented in Figures 5.4.2-5.4.6.

Looking at countries that stand out when within-testlet dependency is of concern, the elaborations listed below are only speculative and would require an in-depth knowledge regarding all PISA's items and educational systems of all countries. Nonetheless, the consistency of this study's results gives justification to some explanations being proposed. Greece (GRC) is frequently featured as an outlier in mathematics (Table 6.2.2), reading (Table 6.2.5) as well as science (Table 6.2.8). Out of all the 24 OECD countries that were investigated, Greece is at the low end of

performance rankings. Perhaps the outlying results are an artefact of this. Another stipulated possibility could be due to students from Greece being taught by specific methods permitting to acquire skills to take advantage of common stimuli that link items within the same testlets. Another option to propose could be 'academic cheating', which would express itself as LID (Zimmermann, Klusmann, & Hampe, 2016). Davis, Drinan, and Bertram Gallant (2009) point out in supplementary Q&A with the authors (Davis, Drinan, & Bertram Gallant, 2017) that

> The research has not shown extensive differences between countries, except that perhaps in countries rife with corruption, there will more cheating (and more serious cheating) in educational systems. Also, in countries where the gap between the "have's" and the "have nots" is greater, cheating is much more likely as well.

Greece is listed in the Corruption Perceptions Index (Transparency International, 2017) as the most corrupted out of all the 24 OECD countries used in this research. Furthermore, research from an educational setting (Dimitriadou, Gakoudi, Kalaitzidou-Leontaki, & Kousaridis, 2012) also acknowledges academic dishonesty in Greece is observed, although it does not offer a cross-national comparison. A final possibility for the observed results in Table 6.2.2, Table 6.2.5 and Table 6.2.8 may be due to either an organised or accidental utilisation of practice for the PISA assessment. The inclusion of small countries such as Luxembourg (LUX) and Iceland (ISL) in the tables listed above, in which the same school would participate in each PISA wave, may be indicative of at least teachers' familiarity with the study.

With regard to the outlying countries when positive LID between-testlets was investigated as reported in Tables 6.2.1, 6.2.4 and 6.2.7, the across domains common conclusion seems to indicate higher levels of this type of LID in highly performing countries such as Finland (FIN), Korea (KOR) and Japan (JPN) with the exception of Ireland (IRL) in PISA 2012. This trend seems to be more visible in the case of mathematics. Perhaps students from these countries can more flexibly take advantage of similarities of questions from across different testlets in regard to the curriculum or item format and characteristics.

With regard to countries that indicated inflated levels of negative LID, Greece was featured most frequently. This could be an artefact of positive LID being observed in Greece as suggested by Habing and Roussos (2003). It is also possible that students in Greece implemented selective time and effort allocation that could indicate negative LID (Yen, 1993) resigning to taking on questions which appeared to be difficult.

6.3 Comparing international and national level LID and seeking differential testlet functioning

As was elaborated in the methodology (see section 3.4.1) and is acknowledged in limitations (see section 7.2.1), comparing the percentages of residual correlations exceeding cut point of 0.1 among countries or against international calibration samples is challenging due to considerably varying sample sizes of student cohorts used in the PISA studies by different countries and the influence of sample size on the expected distribution of RCs mentioned in a recent publication by Christensen et al. (2017). Consequently, research question RQ_3C is addressed by using fractional ranks of residual correations expressed as a percentage. The values of the fractional ranks approximating 100% are indicative of highly ranked positive residual correlations while fractional ranks close to 0% indicate high negative residual correlations. High positive residual correlations are used in this study as indicative of positive dependency among items while negative RCs suggest negative LID. While fractional ranks were obtained independently for each PISA wave, cognitive domain and country, they are reported in subsequent tables together for the sake of cross-national comparison and reference to international results.

This section is subdivided into six sub-sections looking at each cognitive domain separated by the within-testlet or between the testlets location of items' pairs. The tables in this section show comparisons of ranks for all 24 investigated OECD countries along with the ranks of residual correlations from the international calibration samples.

The organisation of tables within each cognitive domain follows largely the same pattern.

The within-testlet perspective is reported within each cognitive domain as the first subsection. Pairs of items indicative of positive within-testlet LID with high cross-wave and cross-national consistency are presented in the first table. The second and third tables show the within-testlet positive LID. The second table represents testlets for which most countries reveal high positive RCs but a few countries do not follow, while the third table focuses on testlets with LID only for few nations. These two descriptive tables are designed to show the possibility of differential testlet functioning. The next table presents testlets with consistent lack of positive within-testlet LID confirmed cross-nationally and across PISA waves. These types of tables are reported to highlight the testlets that firmly adhere to local item independence in regard to the presence of a common stimulus.

The between-testlet perspective is reported for each cognitive domain in the second subsection. Once again the results are organised in the same order for all domains starting with between-testlet

item pairs featuring some cross-national consistency in positive dependency followed by evaluating patterns in negative LID. Research questions RQ_3C and RQ_3D are addressed in this subchapter and throughout this section very low or very high fractional ranks will be treated as indicative of negative and positive LID.

6.3.1 Cross-national LID comparison for mathematics and pairs of items within-testlets.

The discussion about the cross-country within-testlet consistency starts with Table 6.3.1 which presents the non-released two-item testlet M496 “Cash Withdrawal”, which was used across four PISA studies. It revealed consistent positive LID in the international calibration and in all selected 24 OECD countries. Other non-released pairs of items from M406 “Running Tracks” and M828 “Carbon Dioxide” that were frequently used in cross-wave linking revealed this consistency. Item doublets from M124 “Walking”, M402 “Internet Relay Chat”, M704 “The Best Car” and M810 “Bicycles” featured in Table 6.3.1 are open for viewing and reproduced in Electronic Appendix for Figure 5.4.3_POSTIVE LID_WITHIN TESTLET - PISA 2003 Mathematics. The plausible reasons for positive LID existence were proposed in section 5.4.1.1. However, the results for these pairs of items also indicate high cross-national consistency in addition to the positive LID reported for the international calibrations as discussed in the previous chapter.

Table 6.3.1 Fractional ranks of RCs expressed as a percentage for pairs of mathematics items that show within-testlet positive LID with high cross-country and cross-wave consistency

PAIR ID	INT	AUS	AUT	BEL	CAN	CHE	CZE	DEU	DNK	ESP	FIN	GBR	GRC	HUN	IRL	ISL	ITA	JPN	KOR	LUX	NOR	NZL	POL	PRT	SWE
2000_M_M124Q01_M124Q03T	100	93	99	98	99	95	97	98	99	100	97	98	96	92	99	94	100	97	100	100	100	99	98	99	98
2003_M_M124Q01_M124Q03T	99	99	96	96	100	99	99	97	98	99	98	98	99	94	96	98	99	90	92	98	99	98	99	96	89
2003_M_M402Q01_M402Q02	100	100	100	100	100	100	100	100	99	100	100	100	100	MISS	100	99	100	100	100	99	100	100	100	100	100
2003_M_M406Q01_M406Q02	99	99	99	99	100	100	98	100	96	99	95	99	99	96	99	95	99	98	98	99	98	99	94	94	99
2006_M_M406Q01_M406Q02	98	98	99	99	99	100	98	96	98	99	91	98	95	95	93	97	99	96	93	98	98	95	94	92	93
2009_M_M406Q01_M406Q02	99	98	99	98	99	99	98	96	97	99	98	96	93	98	96	97	99	92	90	98	95	95	96	96	90
2012_M_M406Q01_M406Q02	99	98	99	99	100	99	99	97	98	100	99	97	97	94	93	97	100	97	96	99	96	91	96	95	99
2003_M_M496Q01T_M496Q02	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	99	100	98	99	100	100	100	100	100	100
2006_M_M496Q01T_M496Q02	100	100	100	100	100	100	100	100	100	100	100	100	100	98	99	98	100	99	97	100	99	99	100	100	99
2009_M_M496Q01T_M496Q02	100	100	100	100	100	100	100	100	100	100	99	100	100	100	100	99	100	99	98	100	99	100	100	100	99
2012_M_M496Q01T_M496Q02	100	100	100	100	100	100	100	100	100	100	100	100	99	99	99	99	100	99	99	100	99	100	100	100	99
2003_M_M704Q01T_M704Q02T	100	98	98	99	99	100	94	97	90	100	99	96	97	99	99	97	100	92	98	99	98	92	94	92	99
2003_M_M810Q01T_M810Q02T	100	99	100	100	100	100	99	97	99	100	99	99	99	99	95	100	100	99	100	98	99	99	99	98	99
2006_M_M810Q01T_M810Q02T	99	99	99	99	100	100	99	99	99	100	99	99	97	99	99	97	99	98	98	97	97	97	98	98	98
2003_M_M828Q01_M828Q02	100	100	99	100	100	100	99	97	99	100	100	100	100	100	99	100	100	96	99	98	99	100	100	99	100
2006_M_M828Q01_M828Q02	100	99	98	99	100	99	98	97	99	100	99	99	98	99	94	97	99	98	99	99	98	99	94	98	97
2009_M_M828Q01_M828Q02	98	99	100	98	100	99	99	96	98	100	99	98	98	98	98	98	99	99	98	95	98	96	98	97	99
2012_M_M828Q01_M828Q02	99	98	98	98	100	100	97	95	85	100	99	99	94	98	99	99	99	98	100	96	95	97	84	99	96
2012_M_M909Q02_M909Q03	100	100	99	100	100	100	99	100	100	100	100	100	97	99	99	99	100	99	99	99	99	100	99	100	100

* **MISS** identifies an item which was not administered by some countries as reported in PISA Technical Manual (OECD, 2005b) under "National item deletions."

Table 6.3.2 reports the within-testlet pairs of items suggesting positive LID with exceptions for some countries. Used in PISA 2000 and 2003, the non-released four-item testlet M144 “Cube Painting” appears to be indicative of less prominent positive dependency in Hungary (HUN), Japan (JPN) and Korea (KOR). In a similar manner, the two-item testlet M302 “Car Drive”, for which the items have been released, shows high positive RCs most countries, but not in Finland (FIN), Denmark (DNK) or Austria (AUT) in 2006. Both of the M302 questions were simple with the OECD average percentage of students answering them correctly in PISA 2006 equal to 95% and 81% for the first and second questions, respectively. Item pairs from M413 “Exchange Rate” have high positive RCs in most countries, but not in Iceland (ISL) and Japan (JPN). Pairs M919Q01/M919Q02 from “Zs Fan Merchandise” and M992Q01/M992Q02 from “Spacers” show lower fractional ranks for Germany (DEU) while M954Q01/M954Q02 so does a pair M954Q01/M954Q04 from “Medicine Doses” for Korea (KOR). None of the testlets introduced in PISA 2012 and coded in the M9000 range have been released, so it is not possible to provide soundly-based speculations as to why LID is apparent in most countries, but not in those identified above as exceptions.

Table 6.3.2 Fractional ranks of RCs expressed as a percentage for pairs of mathematics items that show within-testlet positive LID with cross-country and cross-wave consistency for majority of the countries

PAIR ID	INT	AUS	AUT	BEL	CAN	CHE	CZE	DEU	DNK	ESP	FIN	GBR	GRC	HUN	IRL	ISL	ITA	JPN	KOR	LUX	NOR	NZL	POL	PRT	SWE
2000_M_M144Q01T_M144Q02T	99	100	98	98	99	96	99	98	99	92	99	100	87	98	98	99	94	81	84	94	99	100	97	97	98
2003_M_M144Q01T_M144Q02T	100	100	96	89	99	99	99	93	97	98	96	100	95	90	97	90	99	95	96	99	96	94	99	95	96
2000_M_M144Q01T_M144Q03	99	100	97	100	100	98	99	99	100	99	99	100	99	94	100	99	94	87	98	97	97	100	89	99	99
2003_M_M144Q01T_M144Q03	99	100	97	97	100	99	96	95	99	99	97	100	98	71	99	MISS	99	94	99	98	98	98	97	98	95
2000_M_M144Q01T_M144Q04T	98	99	96	94	99	98	96	99	88	95	94	99	99	92	96	50	96	86	97	95	89	99	92	94	97
2003_M_M144Q01T_M144Q04T	99	99	96	88	99	99	98	97	84	96	89	99	88	90	89	84	98	96	90	96	91	97	88	97	76
2000_M_M144Q02T_M144Q03	100	94	100	100	100	100	100	100	100	100	100	99	97	92	99	100	99	90	92	93	99	98	100	100	100
2003_M_M144Q02T_M144Q03	99	100	98	100	100	100	100	100	98	100	100	98	84	84	88	MISS	100	91	92	100	97	90	99	98	99
2000_M_M144Q02T_M144Q04T	99	98	99	99	99	98	95	98	93	98	99	98	93	85	100	93	96	94	97	99	93	90	97	90	99
2003_M_M144Q02T_M144Q04T	98	99	94	99	99	98	99	95	96	99	86	98	89	91	95	98	99	78	87	97	91	79	96	98	76
2000_M_M144Q03_M144Q04T	100	100	100	100	100	99	100	100	100	100	100	100	100	99	99	100	100	100	98	100	100	100	100	100	99
2003_M_M144Q03_M144Q04T	100	100	99	100	100	100	100	99	100	100	100	100	96	99	100	MISS	100	99	99	100	100	99	100	100	98
2003_M_M302Q01T_M302Q02	100	100	100	100	100	100	99	99	87	100	38	100	98	100	99	99	100	100	100	100	100	96	100	100	98
2006_M_M302Q01T_M302Q02	100	100	54	100	100	100	99	99	94	100	39	99	99	100	94	98	100	100	100	98	99	100	100	89	99
2003_M_M413Q01_M413Q02	100	100	100	100	100	100	100	99	100	99	100	100	99	99	100	96	99	95	99	99	100	97	97	96	100
2003_M_M413Q01_M413Q03T	98	92	99	97	99	98	94	92	97	99	93	92	97	80	88	87	96	78	88	94	96	88	97	90	99
2003_M_M413Q02_M413Q03T	100	99	98	99	100	99	98	97	99	99	100	99	96	99	99	86	99	92	97	99	98	93	99	97	99
2012_M_M919Q01_M919Q02	98	99	96	99	99	98	96	75	97	100	99	96	96	96	95	96	100	98	97	99	96	99	98	94	99
2012_M_M954Q01_M954Q02	100	99	97	99	100	100	99	98	99	99	100	100	97	98	94	95	98	98	60	99	99	98	99	96	99
2012_M_M954Q01_M954Q04	99	99	99	99	99	100	96	98	98	99	99	98	94	94	98	96	99	92	88	99	99	90	95	97	97
2012_M_M992Q01_M992Q02	100	99	99	100	100	97	96	78	99	100	99	99	100	97	96	90	100	99	100	100	99	94	100	99	95

* **MISS** identifies an item which was not administered by some countries as reported in PISA Technical Manual (OECD, 2005b) under "National item deletions."

For some within-testlet item pairs, no positive LID is indicated in the majority of countries, but in a small number of countries (less than one quarter) very high ranks of residual correlation indicative of positive LID are reported in Table 6.3.3. Positive LID in pairs M155Q01/M155Q02 and M155Q01/M155Q04 could be argued for approximately a quarter of the countries; it is not indicated in Ireland (IRL), Iceland (ISL), Denmark (DNK), Finland (FIN), Japan (JPN) and Korea (KOR) to mention a few. While this testlet is not released, it is known that it tests the mathematical strand of "Statistics, Probability, Data" (DEPP, 2007). Perhaps some curriculum differences are at play with students in Ireland and other nations not being able to take advantage of common testlet introduction or shared between-item underlying knowledge.

Another testlet featured in Table 6.3.3 is the two-item non-released testlet M446 (Thermometer cricket). In data from Austria (AUT), Belgium (BEL), Switzerland (CHE) and Spain (ESP), all reveal substantial positive LID. Positive LID is also apparent in item pairs from testlet M136 (Apples) for Canada (CAN), Spain (ESP), Ireland (IRL), Luxembourg (LUX), Norway (NOR) and Sweden (SWE). This testlet is available to the public, and it involves skills in solving quadratic equations. All these examples reveal the possiblity of differential testlet functioning with respect to positive LID and may be reflective of curriculum differences observed at the particular point in time when 15 year old students are invited to participate in the PISA study. Without access to the items themselves and without in-depth cross-national knowledge about mathematics curricula and teaching and learning practices, evidence-based explanations cannot be offered.

However, the specific LID pattern for M446 mentioned above, which holds across four PISA waves with various combinations of mathematical items used, is very interesting due to the mostly 'Germanic language countries' standing out as indicative of positive LID. In the publication (OECD, 2010b, p. 131) question M446Q02 is mentioned as exhibiting differential item functioning across different grades suggesting that the algebraic skills required to answer these questions may not be equally taught at the same grade level to students in different countries. This points to M446 differential testlet functioning being curriculum driven. With PISA being an age-based study, the LID for testlet M446 indicates that in some countries the mathematical skills required to drive item dependency may not be present.

Table 6.3.3 Fractional ranks of RCs expressed as a percentage for pairs of mathematics items that show within-testlet positive LID with cross-country and cross-wave consistency only for few countries

PAIR ID	INT	AUS	AUT	BEL	CAN	CHE	CZE	DEU	DNK	ESP	FIN	GBR	GRC	HUN	IRL	ISL	ITA	JPN	KOR	LUX	NOR	NZL	POL	PRT	SWE
2000_M_M136Q01T_M136Q02T	94	55	93	92	95	91	93	92	85	95	85	73	96	94	95	78	95	81	70	95	95	75	94	94	96
2000_M_M136Q01T_M136Q03T	95	91	92	92	98	96	95	90	84	96	98	87	90	89	97	91	90	84	93	98	98	80	93	92	96
2000_M_M155Q01_M155Q02T	96	97	92	98	98	97	97	95	93	96	94	94	90	98	88	75	88	MISS	99	80	87	93	96	88	90
2003_M_M155Q01_M155Q02T	99	99	95	99	99	96	99	97	94	99	86	95	99	98	84	93	99	87	MISS	96	94	96	95	91	93
2006_M_M155Q01_M155Q02T	94	98	91	96	98	95	91	88	82	94	79	94	86	87	77	74	97	77	88	88	95	88	94	79	82
2009_M_M155Q01_M155Q02D	98	98	97	96	99	97	97	97	94	97	53	96	98	91	61	92	97	95	95	91	89	81	94	80	86
2012_M_M155Q01_M155Q02D	98	97	91	97	99	97	96	93	88	98	95	97	97	81	89	87	99	96	96	90	92	90	85	87	81
2000_M_M155Q01_M155Q04T	96	98	95	97	97	97	96	97	96	97	98	97	95	96	91	89	99	MISS	86	85	96	83	80	91	92
2003_M_M155Q01_M155Q04T	98	99	85	98	98	98	95	94	84	98	96	89	98	94	91	99	99	93	MISS	94	95	75	95	75	98
2006_M_M155Q01_M155Q04T	94	91	76	98	96	96	95	83	89	98	77	95	89	86	86	94	97	32	74	68	88	85	87	91	92
2009_M_M155Q01_M155Q04T	97	95	91	97	98	97	96	96	90	97	90	96	99	73	85	85	99	85	84	94	92	92	94	88	93
2012_M_M155Q01_M155Q04T	95	96	94	86	96	98	75	79	93	97	98	96	97	96	88	80	99	95	91	94	94	76	82	92	94
2003_M_M446Q01_M446Q02	97	94	99	99	99	98	80	97	68	99	29	81	99	97	98	70	95	77	63	68	24	74	97	58	87
2006_M_M446Q01_M446Q02	93	96	98	99	88	98	52	97	62	99	79	30	55	52	86	61	74	56	98	88	6	47	88	40	84
2009_M_M446Q01_M446Q02	77	86	99	98	96	98	24	94	87	99	36	45	95	56	34	39	99	36	93	94	20	97	80	46	36
2012_M_M446Q01_M446Q02	87	54	87	99	96	99	65	78	34	99	82	53	35	74	98	91	81	47	77	93	37	35	73	62	98

* **MISS** identifies an item which was not administered by some countries as reported in PISA Technical Manuals (Adams & Wu, 2002; OECD, 2005b) under "National item deletions."

Throughout the entire investigation in Chapter 5, there was only one pair of within-testlet items indicative of negative dependency. Table 6.3.4 represents this pair of items from non-released two-item testlet M998 “Bike Rental” that indicates within-testlet negative LID (close to 0 percent rank) for the international calibration and for some of the national samples. It would be worthwhile to review both of these questions to determine whether one of the items was potentially confusing or perhaps even non-appealing for students. While negative LID was found when the difficulties of the pair of items were large (see section 5.4.2.3), the difference in difficulties for the M988 pair is similar to that found in many other item pairs.

The final Table 6.3.5, illustrating within-testlet item pairs, contains the two-item testlet M411 “Diving”, the three-item M995 “Revolving Door” and the four item M982 “Employment Data”. None of these testlets show fractional ranks close to the extremes suggesting that dependency was unlikely to be present. Reviewing the questions from M995 which was released to the public (National Center for Education Statistics, 2016a) gives some justification to this, as all three questions have their own prompts with little need to refer to the graph and introduction at the start of the testlet. These questions also require different knowledge for example, the number of degrees in a circle, calculation of circle circumference and algebraic calculations. These three testlets are examples of a desirable approach for constructing multiple item cognitive tasks, intending to reflect real life multifaceted problems that students are likely to encounter in the workforce.

Table 6.3.4 Fractional ranks of RCs expressed as a percentage for a single pair of mathematics items that showed within-testlet negative LID for international calibration data along with cross-national comparison

PAIR ID	INT	AUS	AUT	BEL	CAN	CHE	CZE	DEU	DNK	ESP	FIN	GBR	GRC	HUN	IRL	ISL	ITA	JPN	KOR	LUX	NOR	NZL	POL	PRT	SWE
2012_M_M998Q02_M998Q04T	5	3	11	24	2	4	8	21	24	14	8	6	29	32	13	67	4	24	7	13	11	8	24	26	22

Table 6.3.5 Fractional ranks of RCs expressed as a percentage for pairs of mathematics items that consistently do not indicate any within-testlet positive LID

PAIR ID	INT	AUS	AUT	BEL	CAN	CHE	CZE	DEU	DNK	ESP	FIN	GBR	GRC	HUN	IRL	ISL	ITA	JPN	KOR	LUX	NOR	NZL	POL	PRT	SWE
2003_M_M411Q01_M411Q02	67	77	90	96	94	79	50	69	63	69	88	59	84	61	85	37	93	77	84	77	62	59	63	65	67
2006_M_M411Q01_M411Q02	65	41	71	90	85	57	40	49	90	70	53	73	80	35	68	82	89	86	31	75	62	47	54	84	72
2009_M_M411Q01_M411Q02	90	73	64	78	89	88	89	71	87	88	37	50	75	36	62	48	97	81	85	76	80	85	55	79	52
2012_M_M411Q01_M411Q02	70	47	92	70	76	90	69	74	73	71	81	55	72	52	43	59	85	63	55	64	81	69	67	57	63
2012_M_M955Q01_M955Q02	61	69	88	97	93	85	84	65	98	94	95	86	85	86	69	80	96	64	68	83	87	72	74	84	62
2012_M_M955Q01_M955Q03	29	4	48	72	24	42	15	9	21	43	27	32	48	53	34	52	17	15	28	24	56	16	22	41	52
2012_M_M955Q02_M955Q03	74	75	53	48	82	86	33	43	79	60	45	62	68	27	70	82	76	49	60	36	78	54	63	22	83
2012_M_M982Q01_M982Q04	62	39	16	36	77	66	63	16	92	73	76	55	67	37	49	66	32	79	43	25	48	36	77	65	98
2012_M_M982Q02_M982Q03T	53	56	71	63	70	70	57	61	66	94	57	86	94	62	54	58	94	87	33	69	66	42	44	50	68
2012_M_M982Q02_M982Q04	40	50	86	61	70	56	81	81	78	83	60	34	49	49	55	39	83	81	40	52	63	28	56	54	57
2012_M_M982Q03T_M982Q04	87	88	62	68	96	81	55	89	96	76	91	78	86	57	45	94	88	96	80	69	93	78	85	51	93

6.3.2 Cross-national LID comparison for mathematics and pairs of between-testlet items

The same percentage rank which was applied independently to the RCs from national and international samples is reported in Table 6.3.6, showing the selected pairs of items from different testlets. This table indicates that for some pairs of items from different testlets positive LID may be present despite items being used in different PISA waves with different cohorts of students. For example, the item pair M408Q01T from "Lotteries" and M423Q01 from "Tossing Coins," while non-released, have testlet titles that suggest a common mathematical strand that is likely to be driving the observed positive LID. This is confirmed in a number of publications (DEPP, 2007; OECD, 2014b) that state that both items belong to the "Statistics, probability and data" mathematical strand. These results are also confirmed in logistic regression models (see section 5.4.2.1) when after controlling for items' pair location, the odds of finding positive LID are five times higher for the pair from "Statistics, probability and data" strand as compared to when both items are of "Number" nature. The presence of positive LID for this pair is constant across all countries.

The remaining items listed in Table 6.3.6 have reported LID ranks close to 100 for a large proportion of countries. Item M408Q01T in pairing with M421Q02T, which is also an "Uncertainty and data" item, indicate the possibility of positive LID in some countries. Pairs of between-testlets items M421Q01/M468Q01T and M509Q01/M710Q01 also originate from the same mathematical competency, but these four items were released to the public. Both M509Q01 from "Earthquake" and M710Q01 from "Forecast of Rain" ask students to select the statement which correctly reflects a prompt giving the percent of chance of an event. M421Q01 from "Height" and M468Q01T from "Science Tests" require the knowledge of how to calculate an average. The remaining last eight items' pairs reported in the table do not come from the "Uncertainty and data" competency, yet are indicative for positive LID in the majority of countries. The item pair containing M145Q01T from "Cubes" and M555Q02T from "Number Cubes", both of which are released, is reported in Table 6.3.6 and its positive dependency is related to the rule of opposite sides of dice adding to seven. As none of the remaining item pairs shown in Table 6.3.6 were released to the public and their item characteristics are unknown for M900s items, the pursuing of plausible reasons for positive LID is available only for PISA study custodians with access to confidentialised items.

Table 6.3.6 Fractional ranks of RCs expressed as a percentage for pairs of mathematics items that indicate between-testlet positive LID with at some degree of cross-country and cross-wave consistency

PAIR ID	INT	AUS	AUT	BEL	CAN	CHE	CZE	DEU	DNK	ESP	FIN	GBR	GRC	HUN	IRL	ISL	ITA	JPN	KOR	LUX	NOR	NZL	POL	PRT	SWE
2003_M_M145Q01T_M555Q02T	99	99	99	99	100	100	99	96	98	99	99	99	96	92	95	97	98	99	98	98	99	98	98	85	99
2000_M_M159Q05_M192Q01T	99	98	98	90	98	99	98	93	84	98	96	99	64	100	95	99	93	99	50	83	99	75	86	70	91
2003_M_M408Q01T_M421Q02T	99	99	98	96	84	94	96	82	94	99	81	96	92	98	96	82	98	96	100	83	96	62	99	99	97
2006_M_M408Q01T_M421Q02T	97	96	32	84	97	92	99	99	83	93	97	74	98	93	94	54	99	92	34	98	72	93	21	80	77
2003_M_M408Q01T_M423Q01	100	100	99	100	100	100	100	99	99	100	100	100	98	100	100	99	100	100	100	98	99	100	100	100	100
2006_M_M408Q01T_M423Q01	100	100	100	100	100	100	100	100	100	100	100	100	96	98	100	99	100	99	100	99	100	100	100	100	100
2009_M_M408Q01T_M423Q01	100	100	100	100	100	100	100	100	MISS	100	100	100	97	100	100	100	100	100	99	100	100	100	100	100	100
2012_M_M408Q01T_M423Q01	100	100	100	100	100	100	100	100	100	100	100	100	98	100	100	99	100	100	100	100	100	97	100	100	100
2003_M_M421Q01_M468Q01T	100	100	100	96	100	95	99	98	99	97	95	100	85	99	99	97	99	94	75	56	99	100	99	99	100
2003_M_M509Q01_M710Q01	99	86	95	97	99	92	95	98	94	94	99	91	82	96	98	11	46	97	85	92	88	99	100	53	100
2012_M_M564Q01_M943Q02	99	99	13	95	51	99	91	100	99	88	100	80	91	100	100	41	99	97	92	90	52	99	87	97	72
2012_M_M909Q01_M955Q01	96	99	86	88	99	86	91	73	58	94	84	88	90	80	90	95	97	95	87	92	86	96	97	97	97
2012_M_M919Q01_M982Q01	99	99	99	98	100	100	98	86	98	92	93	97	43	82	96	99	100	90	99	95	90	73	78	93	37
2012_M_M943Q02_M949Q03	99	90	80	74	98	93	90	96	96	95	99	96	99	90	100	98	98	78	41	64	92	99	72	48	96
2012_M_M992Q03_M998Q04T	100	100	99	66	99	95	73	100	98	100	96	99	95	91	50	51	95	17	98	100	100	100	96	70	48
2012_M_M995Q02_M998Q04T	96	100	99	99	98	100	90	18	100	67	89	99	99	100	99	66	99	61	94	100	92	59	25	100	94

* **MISS** identifies an item which was not administered by some countries as reported in PISA Technical Manual (OECD, 2012) under "National item deletions."

Given that positive LID appears to be present with some cross-national consistency among item pairs within the same testlets and also between different testlets, it is shown mathematically that negative LID must be present in some form as suggested by Habing and Roussos (2003) and van Rijn and Rijmen (2015). However in search for other reasons for negative LID, which as suggested by Yen (1993) may meaningfully be driven by selective effort and time allocation, item pairs with low fractional ranks pointing to negative RCs were identified.

Table 6.3.7 shows item pairs with negative LID and is separated into two subsections. The first six rows present item pairs which both come from two testlets, reporting high within positive LID among their items. The last three rows include item doublets for which only one question originates from a testlet, previously proven to have positive within-testlet LID. This information may point to negative LID being a mathematical artefact of positive LID as per suggestions by Habing and Roussos (2003). On the other hand all questions in this table show very high differences between items' difficulties as expressed by differences in the percentage of students that answered correctly[50] (Minimum difficulty difference=59%, Max difficulty difference = 89%). This is in accordance with the results that were obtained in multilevel logistic regression models explaining negative LID presented in Table 5.4.4. This fact may point towards Yen (1993) explanations for negative LID.

Furthermore, identifying where these pairs were located within booklets offers some interesting regularities. Items are nested under the testlets, which in turn are allocated to a half an hour's worth of testing time clusters (called blocks in PISA 2000). Clusters are then assigned to the booklets following the rotation design reported in the technical manual for each wave.

While the rotational design assures that each cluster is located in a different time slot throughout the duration of the two-hour long test, the temporal order of the clusters within each booklet is fixed. This may create conditions suitable for negative LID as suggested by Yen (1993). So, for example, students were exposed to both PISA 2003 items from pairs M124Q01 and M302Q01T only in Booklet 7. This booklet had the very straightforward item M302Q01T (95% correct) from cluster M7 in the first half an hour of testing, while the more difficult item M124Q01 (36% correct) from cluster M3 which was located in the last half an hour of the two-hour test. In PISA 2003, pairs M302Q01T/M406Q01 and M302Q01T/M406Q02 were located in a booklet for which the first 90 minutes of testing time was dedicated to mathematical items. In PISA 2006 the same two pairs of questions were together only in Booklet 3 which had these two items located in clusters in the

[50] Items' difficulties were extracted from the item classification tables in the PISA Technical Manuals and represent "International percent of correct responses".

second half of the test. In PISA 2003, the pair M302Q01T and M446Q02 was located in Booklet 7 opening with 90 minutes of mathematics, with the more difficultM446Q02 item (7% correct) printed in the last quarter of the booklet. A pair of items M909Q01 and M995Q02 introduced in PISA 2012 and listed in Table 6.3.7 were together only in Booklet 1, with the easier item M909Q01 (89% correct) presented to students in the first quarter of testing time while the more challenging item M995Q02 (3% correct) was placed in the third quarter. This paragraph indicates/suggests that it would be of interest, mainly for logistic regression models, to quantify the relative positioning of the items' pairs in the perspective of a 2 hour long PISA assessment, as it could be proved to be an important predictor of negative dependency. This observation was not anticipated at the data organising stage and is acknowledged as a limitation in section 7.2.1.

The final table (Table 6.3.8) refering to mathematics items shows a number of item doublets reporting a very peculiar pattern of cross national inconsistency, indicating positive or negative LID. All but two pairs of items (M302Q01T_M438Q01 and M302Q01T_M438Q02) presented in this table have a difficulty difference exceeding 60%. Frequently featured items M302Q01T from released testlet "Car Drive" was very easy for students with 95% of PISA 2003 international sample students answering this item correctly. The same can be said about M800Q01 from "Computer Game" (92% students got this item right) and M413Q01 from "Exchange Rate" (80% correct responses).

Table 6.3.7 Fractional ranks of RCs expressed as a percentage for pairs of mathematics items that show between-testlet negative LID with some degree of cross-country and cross-wave consistency

PAIR ID	INT	AUS	AUT	BEL	CAN	CHE	CZE	DEU	DNK	ESP	FIN	GBR	GRC	HUN	IRL	ISL	ITA	JPN	KOR	LUX	NOR	NZL	POL	PRT	SWE
2003_M_M124Q01_M302Q01T	0	0	0	2	2	1	7	96	0	63	37	89	68	4	1	1	0	8	3	1	42	7	0	0	12
2000_M_M136Q03T_M159Q02	0	1	9	17	5	10	83	22	25	66	2	5	92	6	5	11	36	2	2	54	5	46	1	1	3
2003_M_M302Q01T_M406Q01	45	11	18	91	0	32	3	20	0	1	40	0	94	1	2	95	2	49	27	5	47	0	6	1	29
2006_M_M302Q01T_M406Q01	1	2	1	0	14	0	0	0	3	5	1	2	1	3	2	55	0	3	75	1	0	7	1	1	16
2003_M_M302Q01T_M406Q02	86	1	7	93	0	2	2	1	0	1	5	13	55	2	0	90	2	21	2	1	22	2	10	1	0
2006_M_M302Q01T_M406Q02	0	0	0	1	28	0	7	1	1	1	0	5	2	1	1	35	0	2	16	0	11	1	6	0	1
2003_M_M302Q01T_M446Q02	6	0	0	29	12	0	0	1	0	0	2	1	0	0	0	5	0	1	4	0	7	0	0	0	0
2006_M_M302Q01T_M446Q02	14	0	0	0	18	7	0	0	0	0	1	0	1	0	0	0	0	11	38	0	0	1	0	0	0
2012_M_M909Q01_M995Q02	3	0	14	2	0	0	12	4	1	3	28	0	2	0	3	8	4	2	6	0	7	0	1	2	1

Table 6.3.8 Fractional ranks of RCs expressed as a percentage for pairs of mathematics items that show inconsistent pattern of positive or negative LID for different nations

PAIR ID	INT	AUS	AUT	BEL	CAN	CHE	CZE	DEU	DNK	ESP	FIN	GBR	GRC	HUN	IRL	ISL	ITA	JPN	KOR	LUX	NOR	NZL	POL	PRT	SWE
2003_M_M124Q03T_M302Q01T	0	20	0	0	31	19	100	86	0	79	29	100	50	1	100	20	0	1	1	100	33	100	0	100	100
2003_M_M155Q03T_M302Q01T	1	3	1	0	99	100	100	49	100	1	6	59	39	0	7	10	100	3	100	1	79	100	100	0	100
2006_M_M155Q03T_M302Q01T	3	32	100	100	12	28	1	1	100	4	100	91	2	100	19	14	100	0	14	100	100	97	0	100	19
2003_M_M302Q01T_M438Q01	0	41	8	26	0	0	1	77	100	56	98	1	41	21	95	0	96	0	0	22	14	2	85	22	94
2003_M_M302Q01T_M438Q02	0	0	65	0	1	0	0	91	0	3	29	99	28	6	22	1	0	0	0	1	1	10	0	1	72
2003_M_M302Q01T_M462Q01T	14	0	0	6	14	3	100	100	100	100	100	100	54	34	100	1	92	0	0	100	100	100	100	100	4
2006_M_M302Q01T_M462Q01T	3	0	100	100	32	16	100	1	100	80	100	100	1	100	100	100	100	1	18	100	100	100	0	100	100
2003_M_M413Q01_M421Q02T	7	1	1	1	10	0	0	57	57	5	11	2	27	1	0	0	1	21	3	15	1	3	95	0	98
2012_M_M800Q01_M995Q02	99	25	99	91	75	94	3	26	0	99	78	4	15	0	2	1	62	29	0	92	5	79	2	69	1

6.3.3 Cross-national LID comparison for reading and pairs of items within-testlets

This section of the chapter focuses on reading. It follows the pattern established in reviewing mathematics item pairs: within-testlet item pairs revealing consistently high LID between countries are considered first; after which item pairs displaying positive LID but inconsistently between countries are discussed; and finally, item pairs revealing negative LID are discussed.

Within-testlet item pairs indicating consistent positive LID for the international calibration as well as all 24 OECD countries are shown in Table 6.3.9. The released item pair R040Q03A and R040Q03B from R040 "Lake Chad" shows LID. The first question in this pair asks about locating the year in which the graph starts while the second item asks students why that starting point was chosen. While pair R040Q03A and R040Q03B was only used in the first wave of PISA 2000, pair R104Q01 and R104Q05 from R104 "Telephone" revealed consistent across-nation positive LID in all four PISA waves for which it was used. While this pair was not released to the public, both items are of a non-continuous text format (OECD, 2012) and questions from this testlet relate to some form of a table (Soussi et al., 2004). None of the other item pairs featured in Table 6.3.9 are released to the public, but the table confirms cross-wave and cross-national consistency in pointing to positive within-testlet LID. All these items reported in the table below were featured in Section 5.4.1.3 with plausible reasons for LID being offered. Positive dependency for these items may be more likely to be related to the need for referring to an introductory text or graph and therefore logically should be seen to be consistent across all 24 countries.

Table 6.3.9 Fractional ranks of RCs expressed as a percentage for pairs of reading items that show within-testlet positive LID with high cross-country and cross-wave consistency

PAIR ID	INT	AUS	AUT	BEL	CAN	CHE	CZE	DEU	DNK	ESP	FIN	GBR	GRC	HUN	IRL	ISL	ITA	JPN	KOR	LUX	NOR	NZL	POL	PRT	SWE
2000_R_R040Q03A_R040Q03B	100	100	100	100	100	100	100	100	100	100	100	100	100	100	99	100	100	100	100	99	100	100	99	100	99
2000_R_R083Q02_R083Q03	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
2009_R_R083Q02_R083Q03	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
2000_R_R104Q01_R104Q05	100	98	99	100	100	100	100	100	97	98	100	100	100	100	98	100	100	99	100	100	100	99	100	99	99
2003_R_R104Q01_R104Q05	99	99	99	100	100	100	100	99	99	99	98	99	99	99	99	97	98	100	99	98	100	99	99	99	99
2006_R_R104Q01_R104Q05	99	99	99	99	99	100	99	100	100	99	98	99	94	99	99	99	99	99	98	98	97	99	100	100	99
2009_R_R104Q01_R104Q05	100	100	100	100	100	100	100	99	100	100	100	100	100	100	MISS	97	100	97	99	100	100	99	100	100	99
2000_R_R219Q01E_R219Q01T	100	100	100	100	100	100	98	100	100	100	100	100	100	100	100	100	MISS	100	99	100	100	100	100	100	100
2003_R_R219Q01E_R219Q01T	100	100	100	100	100	100	89	100	100	100	100	100	100	99	100	100	100	100	100	99	100	100	98	99	100
2006_R_R219Q01E_R219Q01T	100	100	100	100	100	100	96	100	99	100	100	100	100	100	99	100	100	100	99	98	99	100	97	98	100
2000_R_R220Q05_R220Q06	100	99	100	100	100	100	100	100	100	100	100	100	100	100	100	98	100	81	93	99	100	99	96	100	99
2003_R_R220Q05_R220Q06	100	100	100	99	99	99	99	99	99	98	100	100	100	100	100	80	100	94	98	99	98	100	98	100	97
2006_R_R220Q05_R220Q06	100	100	100	100	100	99	99	99	100	100	100	100	100	99	100	99	100	97	92	100	99	100	100	100	99
2009_R_R220Q05_R220Q06	100	100	100	100	100	100	99	100	100	100	100	100	100	100	100	99	100	96	96	100	99	99	98	MISS	99
2009_R_R404Q10A_R404Q10B	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	99	100	100
2012_R_R404Q10A_R404Q10B	99	100	99	100	100	100	97	100	100	100	100	99	99	97	99	100	100	97	100	99	100	99	98	97	99
2009_R_R406Q01_R406Q05	100	98	99	100	100	100	100	100	100	100	96	99	100	99	92	98	100	100	99	100	100	98	100	100	100
2012_R_R406Q01_R406Q05	100	99	99	99	99	100	99	100	99	100	98	99	100	100	98	100	100	100	98	99	100	99	100	100	99
2009_R_R446Q03_R446Q06	100	100	100	100	100	100	100	100	97	100	100	100	100	100	93	97	100	98	99	92	100	99	98	97	99
2012_R_R446Q03_R446Q06	100	100	92	100	100	99	98	100	87	99	100	100	99	98	100	99	100	99	100	95	100	98	95	99	100

* **MISS** identifies an item which was not administered by some countries as reported in PISA Technical Manuals (Adams & Wu, 2002; OECD, 2012) under "National item deletions."

Table 6.3.10 also reports item pairs for which there was positive LID in the international data set and the majority of the investigated countries. However, for some countries lack of LID in comparison to other nations is visible and this is maintained across multiple PISA waves. For example pairs, R067Q01/R067Q04 and R067Q01/R067Q05 from the non-openly available R067 “Aesop” are in the low range of percentiles ranks only for Japan (JPN) and Korea (KOR). The Korean sample also shows lower fractional RC ranks for a pair of items from R220 “South Pole”. The same indication of differential testlet functioning is present for two pairs of items from R456 “Biscuits” for Hungary (HUN). None of the testlets mentioned above are released to the public. However, it is known that R067 “Aesop” was submitted to the PISA study consortium by Greek authors. It could be assumed that students who attend schools with a European heritage, are more likely to have a greater exposure to Aesop’s literary works through classroom readers and both school and community library catalogues Although it is difficult to determine a plausible reason for Hungary’s lack of dependency in question R456 over two PISA waves, it is interesting to note that this testlet was submitted by Serbia, a southern geographical neighbour to Hungary. The observations that were mentioned above may hint that there are logical explanations for the non-consistent positive LID presence. However, to robustly investigate these suggestions for differential testlet functioning, full access to the relevant cognitive items is needed. Greater knowledge about Japan, Korea and Hungary’s approaches to teaching reading literacy and their language semantics may give justification as to whether their outlying positions in Table 6.3.10 are indeed indicative of a lack of cross-national equivalence for those testlets.

Table 6.3.10 Fractional ranks of RCs expressed as a percentage for pairs of reading items that show within-testlet positive LID with cross-country and cross-wave consistency for majority of the countries

PAIR ID	INT	AUS	AUT	BEL	CAN	CHE	CZE	DEU	DNK	ESP	FIN	GBR	GRC	HUN	IRL	ISL	ITA	JPN	KOR	LUX	NOR	NZL	POL	PRT	SWE
2000_R_R067Q01_R067Q04	100	99	100	99	100	99	99	100	99	99	99	99	89	99	93	83	99	94	67	96	100	96	99	100	99
2003_R_R067Q01_R067Q04	99	99	98	92	98	99	81	99	94	94	81	98	98	94	93	98	98	44	73	91	98	79	98	93	87
2006_R_R067Q01_R067Q04	97	99	99	98	98	99	95	99	95	96	87	98	93	54	97	98	98	88	80	97	85	95	98	99	97
2009_R_R067Q01_R067Q04	98	99	99	97	99	99	97	99	97	99	98	99	100	87	83	95	99	89	66	97	94	89	97	98	100
2000_R_R067Q01_R067Q05	100	99	100	100	100	100	100	100	100	100	100	100	88	96	100	93	100	92	98	100	99	100	100	100	100
2003_R_R067Q01_R067Q05	99	98	99	97	99	98	100	97	88	98	99	99	98	99	89	97	99	85	56	97	96	98	100	95	92
2006_R_R067Q01_R067Q05	98	98	98	99	99	99	98	99	99	98	90	97	96	91	97	99	99	92	89	99	84	97	98	99	98
2009_R_R067Q01_R067Q05	99	99	99	100	99	99	98	98	98	100	89	99	99	98	96	98	100	89	85	98	77	96	97	99	90
2000_R_R067Q04_R067Q05	100	97	98	98	100	99	99	97	99	99	96	100	100	99	100	98	99	100	100	98	100	99	99	97	98
2003_R_R067Q04_R067Q05	99	99	97	99	99	99	96	98	98	99	93	98	99	97	97	92	99	99	100	96	95	99	99	99	93
2006_R_R067Q04_R067Q05	99	97	99	96	99	95	99	96	99	99	99	97	99	98	98	99	98	100	100	95	96	96	99	99	99
2009_R_R067Q04_R067Q05	100	97	96	99	100	97	99	95	97	100	91	99	100	96	86	95	100	98	95	97	95	98	95	95	97
2000_R_R220Q01_R220Q02B	100	98	99	98	100	99	100	99	95	98	98	100	99	99	98	100	98	97	87	97	99	89	100	90	97
2003_R_R220Q01_R220Q02B	98	99	99	98	98	98	99	96	97	97	96	95	97	94	95	92	97	93	92	92	97	94	96	97	96
2006_R_R220Q01_R220Q02B	98	98	94	99	98	98	97	98	MISS	99	96	99	98	96	81	96	97	98	88	95	91	97	99	99	95
2009_R_R220Q01_R220Q02B	99	99	96	98	100	99	98	98	96	100	98	99	100	88	88	98	100	MISS	84	94	98	93	96	98	99
2012_R_R220Q01_R220Q02B	99	100	98	96	99	99	98	98	98	99	99	99	100	95	91	100	99	98	90	97	99	98	97	98	98
2009_R_R456Q01_R456Q02	99	100	87	97	100	100	97	99	98	100	96	100	97	74	99	100	100	94	93	100	93	99	87	99	91
2012_R_R456Q01_R456Q02	99	100	100	94	100	99	95	99	99	100	93	100	96	60	99	97	100	98	99	97	99	99	99	99	99
2009_R_R456Q01_R456Q06	99	100	93	97	100	97	99	62	90	98	98	99	90	81	92	96	99	94	96	97	82	99	91	97	98
2012_R_R456Q01_R456Q06	99	99	95	93	100	97	93	94	99	99	98	100	97	85	93	95	99	93	97	97	97	99	96	96	98
2009_R_R456Q02_R456Q06	100	99	97	100	100	100	100	99	100	100	99	100	99	96	99	95	100	99	88	97	96	99	99	93	98
2012_R_R456Q02_R456Q06	100	100	99	100	100	100	100	99	100	100	100	99	97	99	98	93	100	99	99	99	99	96	99	99	98

* **MISS** identifies an item which was not administered by some countries as reported in PISA Technical Manuals (OECD, 2009b, 2012) under "National item deletions."

On other occasions, some testlets appeared to be positive LID prone only for selected countries, and examples of this may be seen in testlets R228 “Guide” and R245 “Movie Reviews”, listed in Table 6.3.11. It appears that testlet R228 which is not released to the public shows a higher indication of positive LID in Canada (CAN) and Iceland (ISL) and possibly Japan (JPN) and Poland (POL). Similarly, pair R245Q01 and R245Q02 from “Movie Reviews” indicate positive LID for Finland (FIN) and Korea (KOR) which featured in both studies (PISA 2000 and PISA 2009). The same table reports high fractional ranks in Hungary (HUN) and to a lesser extent, Czech Republic (CZE) and Germany (DEU) for R100 “Police”. This testlet is organised in a newspaper article format discussing DNA testing in the text, therefore having a background knowledge of DNA testing may induce item dependency. The released four-item R452 “The Plays the Thing” reveals l higher fractional ranks in Australia (AUS), Canada (CAN), Spain (ESP) and Italy (ITA) (see Table 6.3.11). The testlet starts with the non-typical type of text reporting the script of the theatrical play. It is possible that this kind of prose is utilised in some national curriculums more than others.

Table 6.3.11 Fractional ranks of RCs expressed as a percentage for pairs of reading items that show within-testlet positive LID with cross-country and cross-wave consistency only for few countries

PAIR ID	INT	AUS	AUT	BEL	CAN	CHE	CZE	DEU	DNK	ESP	FIN	GBR	GRC	HUN	IRL	ISL	ITA	JPN	KOR	LUX	NOR	NZL	POL	PRT	SWE
2000_R_R100Q04_R100Q05	88	71	97	84	95	92	78	94	86	94	55	94	87	97	83	94	89	67	53	84	72	84	96	81	73
2000_R_R100Q04_R100Q06	92	70	87	91	96	65	98	98	86	76	92	61	92	99	89	58	94	64	73	78	79	88	97	66	78
2000_R_R100Q04_R100Q07	95	86	65	94	91	81	96	97	77	86	83	66	88	94	87	66	80	98	58	84	90	67	67	92	77
2000_R_R100Q05_R100Q06	96	88	92	39	95	95	98	97	94	66	85	88	93	98	79	80	99	80	77	57	99	97	93	36	99
2000_R_R100Q05_R100Q07	97	98	76	79	87	92	93	81	72	24	95	94	79	96	87	97	93	95	25	73	82	88	78	88	68
2000_R_R100Q06_R100Q07	100	80	96	97	99	98	98	99	96	99	95	75	99	100	93	90	99	80	91	94	99	73	89	89	53
2000_R_R228Q01_R228Q02	96	74	93	72	94	92	67	89	64	60	82	43	52	84	43	93	66	98	93	84	90	45	75	51	92
2000_R_R228Q01_R228Q04	87	90	88	75	99	82	70	83	69	85	97	88	89	82	70	91	59	97	71	79	69	77	97	72	88
2000_R_R228Q02_R228Q04	70	68	80	65	98	66	58	61	55	91	39	91	53	89	50	99	81	80	84	92	86	84	98	41	74
2000_R_R245Q01_R245Q02	75	32	11	63	31	52	19	94	72	84	99	51	40	7	22	66	61	87	98	81	86	10	7	14	41
2009_R_R245Q01_R245Q02	86	90	94	88	92	50	57	80	95	95	98	59	MISS	18	71	MISS	94	64	100	93	91	81	65	66	83
2009_R_R452Q03_R452Q06	97	98	MISS	99	100	99	97	95	93	99	88	97	91	68	98	99	97	95	80	92	96	96	74	94	88
2009_R_R452Q03_R452Q07	99	100	98	93	99	97	88	98	96	99	99	98	88	93	83	96	99	96	92	84	88	98	97	90	96
2009_R_R452Q04_R452Q06	92	97	MISS	84	97	87	93	95	63	98	56	89	95	76	96	74	99	90	81	87	91	74	92	86	73
2009_R_R452Q04_R452Q07	82	94	46	60	78	65	76	78	71	97	55	79	91	81	80	49	92	93	86	88	84	75	88	82	85
2009_R_R452Q06_R452Q07	95	98	MISS	93	97	95	92	96	75	98	88	96	89	76	90	92	97	84	81	97	90	93	63	90	95

* **MISS** identifies an item which was not administered by some countries as reported in PISA Technical Manual (OECD, 2012) under "National item deletions."

Some reading testlets appear dependency neutral (see Table 6.3.12) Testlet R447 “Acne Vulgaris”, which was only used in 2009 and not openly available, showed no within-testlet LID, either in the international sample or in any of the national datasets. This was also the case with testlets R458 “Telecommuting” or R412 “World Languages”. However, it is known that the items in R412 were of different text formats including non-continuous, continuous and mixed. Consequently, it is likely that each item has its own prompt reducing the potential for dependency due to common stimuli. Testlet R458 was released, and the middle question R458Q04 is not at all related to the initial text. The first and last questions require a ‘big picture’ conclusion versus a specific sentence allocation when comparing the two halves of the introductory text.

Table 6.3.12 Fractional ranks of RCs expressed as a percentage for pairs of reading items that consistently do not indicate any within-testlet positive LID

PAIR ID	INT	AUS	AUT	BEL	CAN	CHE	CZE	DEU	DNK	ESP	FIN	GBR	GRC	HUN	IRL	ISL	ITA	JPN	KOR	LUX	NOR	NZL	POL	PRT	SWE
2009_R_R412Q01_R412Q05	89	55	92	92	61	82	89	86	77	93	92	42	96	67	64	70	90	69	43	80	83	56	66	22	86
2012_R_R412Q01_R412Q05	59	54	26	91	67	69	82	82	95	93	67	69	90	90	50	95	94	59	49	80	88	54	80	70	73
2009_R_R412Q01_R412Q06T	41	26	9	60	33	59	8	57	74	32	58	40	60	6	47	30	29	79	29	38	28	29	25	30	46
2012_R_R412Q01_R412Q06T	29	16	19	51	34	46	27	78	58	20	37	49	66	21	67	56	43	67	67	69	53	29	72	42	28
2009_R_R412Q01_R412Q08	73	46	87	93	81	68	80	41	90	69	85	79	80	91	79	22	91	71	95	85	58	82	85	74	61
2012_R_R412Q01_R412Q08	42	62	52	81	67	44	75	37	65	85	95	92	94	59	49	48	86	66	88	65	84	19	57	68	92
2009_R_R412Q05_R412Q06T	67	95	41	96	68	75	88	69	49	50	71	87	71	81	70	70	87	63	81	72	48	64	37	42	52
2012_R_R412Q05_R412Q06T	78	93	82	77	89	87	59	73	67	61	67	60	77	67	60	61	89	77	69	53	74	89	48	65	43
2009_R_R412Q05_R412Q08	70	77	63	74	88	66	84	83	69	93	68	90	93	55	79	45	89	81	79	83	50	81	86	70	53
2012_R_R412Q05_R412Q08	82	81	51	65	77	81	70	88	85	79	90	75	79	55	62	88	81	70	79	79	74	59	70	60	87
2009_R_R412Q06T_R412Q08	75	77	75	82	82	90	73	63	83	76	80	49	59	31	76	88	72	76	75	93	50	71	31	66	82
2012_R_R412Q06T_R412Q08	78	94	74	78	86	86	73	92	77	90	87	77	87	75	42	83	86	86	71	72	92	85	35	58	68
2009_R_R447Q01T_R447Q04	54	69	71	77	71	85	48	62	86	81	68	61	72	46	49	63	64	62	53	58	28	61	54	67	44
2009_R_R447Q01T_R447Q05	51	51	55	68	62	52	69	70	58	56	85	61	47	30	54	57	73	89	48	49	61	62	80	42	83
2009_R_R447Q01T_R447Q06	75	75	59	44	87	80	75	78	81	80	71	78	82	66	39	55	73	39	81	77	67	86	61	73	71
2009_R_R447Q04_R447Q05	60	71	78	58	84	86	76	73	86	80	83	63	54	48	28	67	77	52	41	82	72	79	78	55	55
2009_R_R447Q04_R447Q06	83	81	83	77	82	59	81	89	72	79	51	71	78	80	69	66	90	70	76	83	44	45	54	41	52
2009_R_R447Q05_R447Q06	74	36	39	72	49	69	79	79	52	43	36	89	67	33	70	71	76	81	76	85	52	31	50	68	81
2009_R_R458Q01_R458Q04	54	52	62	54	67	46	52	47	58	58	48	46	48	74	71	71	96	98	47	53	71	50	29	39	60
2009_R_R458Q01_R458Q07	70	88	70	58	81	90	48	85	79	90	55	62	57	65	71	58	71	87	85	77	62	56	52	71	86
2009_R_R458Q04_R458Q07	81	63	74	89	91	87	67	39	69	56	83	75	81	49	42	67	83	66	45	59	61	47	70	50	87

6.3.4 Cross-national LID comparison for reading and pairs of between-testlet items

As was the case in the corresponding section about the mathematics item pairs, reading also has pairs of questions from different testlets which are indicative of positive LID in international and the majority of national datasets. Knowledge about different items' characteristics will be utilised to explain the results in Table 6.3.13, which lists a selection of items introduced in PISA 2009 and reused for the purpose of cross-wave linking again in PISA 2012. Only one pair of items (R099Q04B from "Plan International" and R120Q06 from "Student Opinions") listed in the table came from released testlets. Both of these questions are of an "Open Constructed Response" format, in which students are asked to write their opinions about a topic factoring the information provided in the introductory prompt. High fractional ranks of RCs are present in the international calibrations and only shows for select countries. Because none of the remaining items' doublets reported in Table 6.3.13 are released, the investigation as to whether any of the LID drivers, suggested by Yen (1993), are at play is limited. However, on the basis of various items' characteristics (see section 3.3.2.4 for details as to how they were obtained) items R412Q01 from "World Languages" and R420Q09 from "Children's Futures" were both relatively straight-forward and of a non-continuous text type, i.e. lists, forms, graphs, or diagrams. Both items also share the same "Expository" text type, targeting the "Access and retrieve" aspect of reading while being placed in an "Educational" text situation. On the other hand, items R446Q03 from testlet "Job Vacancy" and R456Q01 from "Biscuits" only share the reading aspect aimed to evaluate "Access and retrieve" skills, but were both very easy, with in excess of 90% of correct responses in the international sample data. Item R446Q03 also indicated dependency with R466Q06 from "Work Right". Both of the questions again shared "Access and retrieve" aspect of reading and were placed in the "Occupational" situation. However, given that the titles of both testlets commonly related to the topic of employment, this may also be offered as a speculative explanation for positive LID indication. Finally, Table 6.3.13 indicates a positive LID between item R455Q03 from "Chocolate and Health" and items R406Q01 and R406Q05 from testlet "Kokeshi Dolls". The only item characteristics that these questions share is being placed in the same "Personal" situation, with a continuous text involved.

However, both testlets in PISA 2009 and PISA 2012 were co-located in the same reading cluster of R5 and PR1, respectively. A pair of questions R101Q02 from Rhinoceros" and R456Q01 from "Biscuits" share a "Continuous" type of text format and "Simple Multiple Choice" item format, but come from different types of text, situation and reading aspects. Questions R412Q06T from testlet "World Languages" and R424Q02T from "Fair Trade" indicate higher fractional ranks for several countries in both PISA studies in which they were used. Both items are of "Complex Multiple

Choice" format with the shared Educational situation and are used to evaluate the "Integrate and interpret" reading aspect. Finally, pair R452Q03 from "The Play's the Thing" and R466Q03T from "Work Right" does not have any item characteristics in common.

Table 6.3.13 Fractional ranks of RCs expressed as a percentage for pairs of reading items that indicate between-testlet positive LID with some degree of cross-country and cross-wave consistency

PAIR ID	INT	AUS	AUT	BEL	CAN	CHE	CZE	DEU	DNK	ESP	FIN	GBR	GRC	HUN	IRL	ISL	ITA	JPN	KOR	LUX	NOR	NZL	POL	PRT	SWE
2000_R_R099Q04B_R120Q06	99	68	50	95	99	89	95	92	67	80	47	65	99	98	78	97	92	100	100	78	71	51	MISS	96	55
2009_R_R101Q02_R456Q01	97	100	82	100	99	100	100	23	94	100	71	99	70	82	99	48	99	100	87	82	87	97	99	98	93
2009_R_R406Q01_R455Q03	97	78	89	94	90	78	97	91	90	99	52	94	MISS	98	65	76	99	97	85	48	95	91	98	95	94
2012_R_R406Q01_R455Q03	93	66	93	85	97	95	96	96	87	99	58	87	97	96	55	81	96	98	98	68	99	84	95	97	95
2009_R_R406Q05_R455Q03	98	98	92	98	96	99	97	96	96	100	84	99	MISS	97	91	70	99	99	99	97	97	94	94	97	94
2012_R_R406Q05_R455Q03	97	97	87	77	95	96	92	98	95	99	95	94	97	95	91	89	98	99	92	93	99	97	98	97	95
2009_R_R412Q01_R420Q09	94	96	96	98	93	84	98	90	96	92	90	98	98	92	95	55	97	94	94	94	98	98	79	85	97
2012_R_R412Q01_R420Q09	92	93	86	86	81	94	94	93	91	84	93	96	93	94	82	89	91	83	85	91	91	92	84	82	61
2009_R_R412Q06T_R424Q02T	99	93	98	85	98	77	94	91	88	82	84	96	100	99	94	96	83	62	68	95	89	69	48	100	84
2012_R_R412Q06T_R424Q02T	96	96	97	79	75	83	98	59	96	98	8	98	87	69	92	63	97	69	88	90	41	95	55	99	98
2009_R_R446Q03_R456Q01	98	97	96	97	95	97	93	99	99	98	60	100	98	73	94	94	98	96	93	76	96	91	49	43	93
2012_R_R446Q03_R456Q01	98	99	90	89	92	94	71	84	97	98	52	100	97	90	88	59	97	99	94	96	95	94	97	100	94
2009_R_R446Q03_R466Q06	98	97	80	100	99	100	96	99	95	98	96	96	94	81	96	91	99	98	93	94	97	92	96	95	88
2012_R_R446Q03_R466Q06	98	95	86	92	99	93	90	96	96	98	99	99	89	90	94	94	98	97	93	95	97	95	97	89	96
2009_R_R452Q03_R466Q03T	100	98	94	100	99	100	99	56	92	99	83	93	92	66	99	99	99	97	98	50	98	32	MISS	45	55

* **MISS** identifies an item which was not administered by some countries as reported in PISA Technical Manuals (Adams & Wu, 2002; OECD, 2012) under "National item deletions."

An indication of negative LID among reading items from different testlets is presented in Table 6.3.14. The pairs of items (R219Q01E/R219Q01T and R220Q05 /R220Q06) which showed high within-testlet positive LID in PISA 2003 and 2006, as reported in Table 6.3.9, also engage in negative LID cross-pairing in these two waves, which can be seen in the first half of Table 6.3.14. This pattern is likely to be a mathematical artefact (Habing & Roussos, 2003; van Rijn & Rijmen, 2015), strongly visible when a limited number of only eight reading testlets is used (see section for 4.2 for more details about using the same testlets across PISA waves) in years when reading was not the targeted assessment domain.

Interestingly, the same mix of R219 and R220 pairs of the item in PISA 2000 does not suggest high negative RCs. Reading was a focus of the first wave of the PISA study in 2000 and therefore involved a much larger number of items. Similar regularity occurs while looking at items from R220 and R404 in Table 6.3.14. The negative between-testlet LID is confirmed in PISA 2012 when at the same time within-testlet positive LID was present in R220 and R404. PISA 2012 was not a reading targeted domain and used a smaller number of items. However when the same testlets are investigated in PISA 2009, also a reading targeted wave with over 100 questions used, the negative LID effect is less pronounced. This regularity discussed here provides extra weight to Habing and Roussos (2003) arguments about the origin of negative dependence, but also gives an explanation for the results in Table 5.2.2 in which the negative LID prevalence appears to be smaller for the literacies which are targeted in the PISA study.

Table 6.3.14 Fractional ranks of RCs expressed as a percentage for pairs of reading items that show between-testlet negative LID with some degree cross-country and cross-wave consistency

PAIR ID	INT	AUS	AUT	BEL	CAN	CHE	CZE	DEU	DNK	ESP	FIN	GBR	GRC	HUN	IRL	ISL	ITA	JPN	KOR	LUX	NOR	NZL	POL	PRT	SWE
2003_R_R219Q01T_R220Q06	3	3	15	14	1	7	3	2	4	7	18	8	25	12	12	27	7	11	23	3	50	56	9	15	2
2003_R_R219Q01T_R220Q05	0	1	12	8	1	1	1	3	1	6	3	2	4	3	18	4	2	15	1	1	11	6	3	1	4
2003_R_R219Q01E_R220Q06	4	3	19	12	1	4	3	6	3	17	13	15	10	9	26	15	4	61	20	3	21	18	12	29	4
2003_R_R219Q01E_R220Q05	1	2	16	1	0	1	2	2	1	12	4	3	3	1	6	6	1	14	6	1	12	10	3	3	3
2006_R_R219Q01T_R220Q06	3	2	13	1	2	5	10	11	5	3	44	5	7	14	11	5	3	16	21	13	16	15	4	10	9
2006_R_R219Q01T_R220Q05	1	0	14	0	1	3	12	6	6	1	13	7	3	11	2	2	4	4	15	10	3	1	0	6	4
2006_R_R219Q01E_R220Q06	2	1	25	1	2	6	8	7	2	3	67	18	14	22	9	7	10	8	19	22	13	25	3	8	9
2006_R_R219Q01E_R220Q05	1	1	22	1	1	11	6	8	9	2	6	9	3	6	7	2	8	1	25	5	10	4	2	16	6
2000_R_R219Q01T_R220Q06	7	37	51	30	4	51	6	29	1	3	57	57	90	39	20	9	16	64	15	85	52	67	48	16	77
2000_R_R219Q01T_R220Q05	52	82	25	11	11	1	1	67	5	9	88	8	73	95	74	60	69	87	72	59	93	95	75	2	85
2000_R_R219Q01E_R220Q06	22	77	46	33	1	19	9	54	5	3	21	78	80	94	41	91	MISS	58	48	31	18	9	58	2	74
2000_R_R219Q01E_R220Q05	58	72	28	15	25	12	17	5	33	2	20	93	33	79	51	51	MISS	59	35	32	95	98	80	2	98
2012_R_R220Q02B_R404Q10B	1	8	19	3	1	1	1	6	4	1	4	2	19	12	14	35	1	4	43	6	2	5	2	2	1
2012_R_R220Q02B_R404Q10A	2	3	13	12	1	3	1	20	4	2	9	5	12	1	13	6	1	3	33	5	21	5	1	1	0
2012_R_R220Q01_R404Q10B	1	16	10	7	4	3	8	2	16	3	17	2	17	14	12	2	3	2	6	5	33	7	2	4	9
2012_R_R220Q01_R404Q10A	1	14	12	19	5	1	4	13	15	2	28	13	25	11	7	2	3	5	10	3	14	17	3	5	1
2009_R_R220Q02B_R404Q10B	47	64	92	89	82	82	30	86	87	98	92	62	20	68	32	87	44	MISS	38	79	72	15	90	63	13
2009_R_R220Q02B_R404Q10A	68	65	30	81	87	75	43	55	87	95	91	79	21	84	30	93	92	MISS	45	61	40	35	73	81	48
2009_R_R220Q01_R404Q10B	51	42	4	94	92	44	19	59	35	83	12	50	21	30	12	28	93	78	15	98	86	36	56	36	33
2009_R_R220Q01_R404Q10A	69	66	21	42	83	19	47	66	63	68	40	52	35	40	43	47	91	71	39	83	43	44	19	86	19

* **MISS** identifies an item which was not administered by some countries as reported in PISA Technical Manuals (Adams & Wu, 2002; OECD, 2012) under "National item deletions."

A different driver for negative LID may be present while looking at low fractional ranks of RCs, which are also reported in Table 6.3.15. The first nine pairs from PISA 2009 came from Booklet B2. In this booklet items R067Q01, R220Q05 and R220Q06 are from cluster R1 which was presented to students in the first half an hour of testing. Items R432Q01, R446Q03, R460Q05, R466Q03T and R466Q06 were placed in the last 30 minutes of testing in the fourth cluster R7. The second half of the Table 6.3.15 points to six item pairs in PISA 2012 that indicate consistent negative LID which all featured together in Booklet number 13. Two "Open Constructed Response" questions from R404 "Sleep" were placed in the first cluster in PISA 2012 Booklet 13 while the matching easier questions R432Q01, R446Q03, R456Q01 were allocated in the last quarter of the testing time. For all six items, the difference between their difficulties (expressed as a percentage of the OECD sample that answered them correctly) was at least 40%. Table 6.3.15 show item pairs for which negative LID may be driven by selective time and effort allocation as suggested by Yen (1993). Item couplings with extreme swings from very high to very low fractional ranks similar to mathematical pairs from Table 6.3.8 were not apparent.

Table 6.3.15 Fractional ranks of RCs expressed as a percentage for pairs of reading items that show between-testlet negative LID with some degree cross-country and cross-wave consistency (cont.)

PAIR ID	INT	AUS	AUT	BEL	CAN	CHE	CZE	DEU	DNK	ESP	FIN	GBR	GRC	HUN	IRL	ISL	ITA	JPN	KOR	LUX	NOR	NZL	POL	PRT	SWE
2009_R_R067Q01_R446Q03	2	0	1	2	0	3	3	0	0	1	4	18	0	89	91	1	2	31	99	4	19	11	31	82	0
2009_R_R067Q01_R466Q03T	0	0	0	100	0	8	14	0	0	0	5	1	22	21	1	73	1	0	38	3	0	23	MISS	5	14
2009_R_R067Q01_R466Q06	1	1	0	1	0	1	2	0	23	1	53	6	2	7	26	6	0	11	96	10	3	8	2	26	2
2009_R_R220Q05_R432Q01	4	3	2	0	1	0	10	2	1	1	9	6	1	0	3	7	1	3	7	0	9	1	5	MISS	18
2009_R_R220Q05_R460Q05	0	56	5	4	2	0	3	52	3	6	54	1	8	2	1	1	3	10	26	20	26	11	76	MISS	3
2009_R_R220Q05_R466Q03T	0	0	4	2	0	0	0	59	6	0	0	1	15	0	17	14	0	37	99	0	2	20	MISS	MISS	37
2009_R_R220Q06_R432Q01	1	7	2	12	0	0	36	5	0	0	16	2	1	1	2	4	0	8	84	4	24	2	5	0	2
2009_R_R220Q06_R460Q05	1	15	10	2	1	5	4	16	4	4	73	2	21	2	8	4	7	3	28	18	13	7	10	0	4
2009_R_R220Q06_R466Q03T	5	82	1	36	14	8	3	0	2	0	1	3	62	0	32	10	2	14	4	0	27	53	MISS	1	2
2012_R_R404Q10A_R432Q01	1	1	39	1	0	7	4	4	11	1	2	3	1	1	3	2	0	6	32	1	11	3	1	3	1
2012_R_R404Q10A_R446Q03	1	3	18	7	0	1	2	13	4	0	1	1	1	25	7	6	1	2	4	1	14	1	16	23	3
2012_R_R404Q10A_R456Q01	1	2	0	16	1	8	20	5	35	2	9	6	7	0	4	0	1	3	5	60	4	0	14	1	1
2012_R_R404Q10B_R432Q01	0	1	21	1	0	4	31	9	10	1	1	3	1	3	22	3	0	9	13	2	12	7	0	2	4
2012_R_R404Q10B_R446Q03	0	1	8	7	0	4	4	5	33	0	1	0	0	20	21	4	0	1	6	1	3	15	1	10	4
2012_R_R404Q10B_R456Q01	0	1	1	2	1	9	4	26	36	1	4	26	22	18	3	1	1	7	11	16	1	1	7	2	9

* **MISS** identifies an item which was not administered by some countries as reported in PISA Technical Manual (OECD, 2012) under "National item deletions."

6.3.5 Cross-national LID comparison for science and pairs of items within-testlets

Investigating LID involving a pair of science items from the same testlets is reported in Tables 6.3.16-6.3.19. Table 6.3.16 shows three pairs of science questions with fractional ranks pointing to positive LID. The results are consistent for national and international datasets, and also when the testlets were used in many PISA studies. Out of the featured testlets in Table 6.3.16, only one (S114 "Greenhouse") has been released, and both questions S114Q03T/S114Q04T require the student to refer to graphs that precede them. Two questions from S326 "Milk" are not available for viewing, but both are of "Simple Multiple Choice" item formats, requiring students to use scientific evidence about a health related topic placed in a personal item context. Dependency for items from S304 "Water" appears to be driven by item chaining, as suggested by the A and B suffixes for this pair of items (304Q03A and 304Q03B).

While Table 6.3.17 also presents within-testlet pairings, which indicate in the majority of countries these items suggested positive LID, some OECD economies are not adhering to the predominant pattern. Items S415Q07T and S415Q08T come from the unreleased testlet called "Solar Panels", but seem to be less LID indicative for Poland (POL), Luxembourg (LUX) and Hungary (HUN) from the perspective of three waves in which these items were used. A speculative reason for the lack of LID in these countries may be due to the students being less exposed to the concepts of solar power or perhaps the competency of "Identifying scientific issues" which both items purported to assess. In the same way, data from Germany (DEU), Austria (AUT), Ireland (IRE) and Iceland (ISL) showed lack of positive LID in three waves for which pair S438Q02/S438Q03 from "Green Parks" was used. This pair also evaluated the competency of "Identifying scientific issues". The last pair of items (S514Q02_S514Q03) in Table 6.3.17 which come from a testlet titled "Development and Disaster", appears to be positive LID indicated in all countries except Korea (KOR). As this item is not released, it is hard to speculate why positive LID is not pronounced for this pair in Korea.

Table 6.3.16 Fractional ranks of RCs expressed as a percentage for pairs of science items that show within-testlet positive LID with high cross-country and cross-wave consistency

PAIR_ID	INT	AUS	AUT	BEL	CAN	CHE	CZE	DEU	DNK	ESP	FIN	GBR	GRC	HUN	IRL	ISL	ITA	JPN	KOR	LUX	NOR	NZL	POL	PRT	SWE
2000_S_S114Q03T_S114Q04T	99	94	97	98	100	93	97	96	96	100	96	97	95	93	91	89	93	96	84	95	96	93	88	97	86
2003_S_S114Q03T_S114Q04T	99	97	95	97	99	96	95	93	93	98	98	99	94	93	94	95	98	92	96	91	90	96	94	98	93
2006_S_S114Q03T_S114Q04T	100	100	95	98	100	100	98	94	97	100	96	100	95	98	93	95	100	99	94	97	94	97	97	96	90
2003_S_S304Q03A_S304Q03B	100	100	98	97	100	99	98	99	98	100	100	99	98	99	98	98	100	99	100	97	98	98	98	98	96
2006_S_S304Q03A_S304Q03B	100	99	97	99	100	100	97	93	99	100	100	100	98	97	97	95	100	99	98	97	98	95	98	95	97
2003_S_S326Q01_S326Q02	100	100	99	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	MISS	100
2006_S_S326Q01_S326Q02	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
2009_S_S326Q01_S326Q02	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
2012_S_S326Q01_S326Q02	100	100	100	100	100	100	100	99	100	100	100	100	100	100	100	99	100	99	100	100	100	100	100	100	100

* **MISS** identifies an item which was not administered by some countries as reported in PISA Technical Manual (OECD, 2005b) under "National item deletions."

Table 6.3.17 Fractional ranks of RCs expressed as a percentage for pairs of science items that show within-testlet positive LID with cross-country and cross-wave consistency for majority of the countries

PAIR_ID	INT	AUS	AUT	BEL	CAN	CHE	CZE	DEU	DNK	ESP	FIN	GBR	GRC	HUN	IRL	ISL	ITA	JPN	KOR	LUX	NOR	NZL	POL	PRT	SWE
2006_S_S415Q07T_S415Q08T	97	89	94	90	98	99	97	91	68	96	93	95	94	77	89	86	98	92	92	57	93	98	75	85	96
2009_S_S415Q07T_S415Q08T	99	97	96	96	99	98	95	89	96	99	99	99	92	96	88	91	97	96	97	97	96	95	93	95	86
2012_S_S415Q07T_S415Q08T	92	99	MISS	91	97	92	94	MISS	95	98	98	96	63	50	91	88	97	92	85	56	96	99	62	83	82
2009_S_S438Q02_S438Q03D	99	98	96	97	100	98	68	54	99	100	98	98	93	97	93	81	98	94	95	83	95	95	95	96	96
2012_S_S438Q02_S438Q03D	98	98	79	99	100	92	95	81	97	99	97	99	98	92	78	59	99	95	96	93	98	95	99	87	94
2006_S_S438Q02_S438Q03T	99	95	72	90	99	93	97	93	95	99	97	97	96	90	84	87	97	95	88	92	99	89	94	93	96
2006_S_S514Q02_S514Q03	100	98	98	98	99	99	100	99	99	100	73	99	99	95	87	97	99	96	79	88	100	95	99	97	96
2009_S_S514Q02_S514Q03	98	100	99	90	100	100	100	96	96	95	100	99	99	90	100	95	98	95	87	90	99	86	97	98	99
2012_S_S514Q02_S514Q03	96	100	95	90	100	97	94	91	96	99	100	98	97	97	98	72	99	97	57	96	98	91	98	95	98

* **MISS** identifies an item which was not administered by some countries as reported in PISA Technical Manual (OECD, 2014b) under "National item deletions."

Table 6.3.18 identifies items' doublets which show positive LID for some countries only. None of these items are openly available. Two questions from S438 "Green Parks" are highlighted showing consistent LID for PISA 2006, 2009 and 2012 in Australia (AUS), Spain (ESP) and Great Britain (GBR). Similarly, two pairs of items from S252 "South Rainea" used in two PISA waves, report high fractional ranks for Switzerland (CHE), Italy (ITA) and possibly Korea (KOR). Interestingly, this testlet was proposed to be included in the PISA items pool by a Korean source and later translated into other languages. The two-item testlet S131 "Good Vibrations" seems to be non-dependent in PISA 2000. However, after this initial PISA wave, the S131Q02/S131Q04 pair had fractional ranks of RCs consistently exceeding rank of 95 percent for Italy (ITA), Greece (GRC), Canada (CAN), Spain (ESP), and Poland (POL). The final item pair from S476 "Heart Surgery" in Table 6.3.18 showed high RC ranks indicating positive LID only for Italy (ITA). Without these items being released, plausible reasons for this positive LID, in a subset of countries, cannot be offered.

Table 6.3.18 Fractional ranks of RCs expressed as a percentage for pairs of science items that show within-testlet positive LID with cross-country and cross-wave consistency only for few countries

PAIR_ID	INT	AUS	AUT	BEL	CAN	CHE	CZE	DEU	DNK	ESP	FIN	GBR	GRC	HUN	IRL	ISL	ITA	JPN	KOR	LUX	NOR	NZL	POL	PRT	SWE
2000_S_S131Q02T_S131Q04T	77	89	77	74	89	44	84	85	98	93	69	95	91	86	55	85	92	86	78	39	74	88	56	72	94
2003_S_S131Q02T_S131Q04T	99	97	92	94	98	84	98	98	93	97	81	95	98	95	81	93	99	96	88	92	81	96	97	98	90
2006_S_S131Q02T_S131Q04T	99	95	71	93	99	96	91	95	79	98	96	96	100	56	85	93	100	100	90	97	93	68	97	92	90
2009_S_S131Q02D_S131Q04D	97	96	91	97	99	92	97	87	97	98	81	89	99	97	87	94	98	93	80	92	88	82	96	91	95
2012_S_S131Q02D_S131Q04D	92	90	91	97	98	80	87	95	91	98	97	89	98	30	85	95	100	98	80	86	88	81	95	90	92
2000_S_S252Q01_S252Q03T	83	63	97	95	90	96	97	94	88	76	94	78	84	80	70	71	95	77	92	70	82	97	60	42	85
2003_S_S252Q01_S252Q03T	84	97	96	83	95	97	94	92	87	94	79	82	97	81	69	91	99	75	92	58	97	87	83	47	77
2000_S_S252Q02_S252Q03T	98	80	93	76	94	98	68	77	89	97	85	92	85	99	62	93	98	84	92	99	95	77	99	69	99
2003_S_S252Q02_S252Q03T	81	94	66	98	98	97	93	92	89	71	89	99	98	92	80	95	97	94	96	68	91	92	80	86	95
2006_S_S438Q01T_S438Q02	88	98	48	95	76	76	72	49	87	97	78	91	90	96	93	82	96	79	90	75	69	59	81	57	46
2009_S_S438Q01T_S438Q02	75	98	61	77	55	94	76	54	79	98	90	99	80	94	92	93	88	86	67	81	85	99	72	76	42
2012_S_S438Q01T_S438Q02	50	95	68	94	94	65	76	70	66	94	89	95	71	76	52	61	88	81	58	60	68	77	89	74	86
2006_S_S476Q01_S476Q03	76	62	61	88	70	89	88	81	75	82	60	35	80	71	21	64	97	91	78	91	59	53	81	80	35
2006_S_S476Q02_S476Q03	96	94	50	88	95	93	96	66	85	88	80	84	84	93	73	85	100	92	94	83	87	87	87	74	93

Finally, Table 6.3.19 show all pairs of items from testlet S498 “Experimental Digestion”, for which no indication of LID was present in any of the three PISA implementations for which this testlet was utilised. While this testlet is confidential, it is known that it consists of three questions, each with a different item format type, starting with “Complex Multiple Choice”, then “Simple Multiple Choice” and finishing with an “Open Response” item. Two other testlets listed in this table, namely the three-item S253 “Ozone” and the two-item S508 “Genetically Modified Crops”, are published. While one of the items from pair S508Q02T/S508Q03 requires a reference to the introductory text, the other does not. Out of three questions in S253, only the middle one requires a reference to the testlet’s main prompt. The first question has its own lengthy introduction, while the third inquires about knowledge not related to the testlet texts. Lack of dependency among S253 and S508 items is explainable.

Table 6.3.19 Fractional ranks of RCs expressed as a percentage for pairs of science items that consistently do not indicate any within-testlet positive LID

PAIR_ID	INT	AUS	AUT	BEL	CAN	CHE	CZE	DEU	DNK	ESP	FIN	GBR	GRC	HUN	IRL	ISL	ITA	JPN	KOR	LUX	NOR	NZL	POL	PRT	SWE
2000_S_S253Q01T_S253Q02	72	90	74	70	92	75	83	34	11	67	86	66	78	85	58	93	41	53	64	82	80	89	42	86	52
2000_S_S253Q01T_S253Q05	79	26	71	74	63	63	55	66	91	69	74	64	84	90	59	37	43	27	63	89	71	21	77	38	39
2000_S_S253Q02_S253Q05	73	56	48	21	34	25	81	47	69	89	21	54	81	50	61	29	64	72	59	62	69	76	89	43	62
2006_S_S498Q02T_S498Q03	63	59	24	37	32	42	40	42	71	58	56	39	71	64	53	68	43	49	70	61	65	41	44	74	26
2009_S_S498Q02T_S498Q03	59	76	41	88	36	83	32	39	41	79	60	86	78	17	20	68	58	73	65	80	69	62	89	74	56
2012_S_S498Q02T_S498Q03	43	58	51	63	52	58	30	58	45	62	27	72	39	55	72	30	75	83	45	64	65	71	72	55	77
2006_S_S498Q02T_S498Q04	64	69	52	43	48	43	77	51	65	42	37	24	62	54	37	47	71	65	69	44	80	45	68	70	46
2009_S_S498Q02T_S498Q04	64	42	85	34	32	60	62	53	32	69	51	44	68	71	20	63	82	88	35	53	73	31	70	86	49
2012_S_S498Q02T_S498Q04	28	55	45	80	36	60	84	42	30	52	65	34	80	48	14	66	83	29	34	70	61	38	52	78	68
2006_S_S498Q03_S498Q04	79	81	53	85	88	55	57	83	53	78	70	81	96	85	86	73	82	76	70	63	57	85	78	75	39
2009_S_S498Q03_S498Q04	48	94	85	51	82	79	47	77	89	79	72	85	83	69	93	68	95	41	59	31	29	76	79	71	38
2012_S_S498Q03_S498Q04	90	90	56	83	79	70	83	55	90	80	82	86	46	74	72	88	69	73	72	83	91	64	78	72	83
2006_S_S508Q02T_S508Q03	91	59	88	80	73	87	96	67	95	95	55	84	86	89	67	81	93	90	84	71	71	74	83	56	75

6.3.6 Cross-national LID comparison for science and pairs of between-testlet items

As was the case in the previous two cognitive domains, there were pairs of items from different testlets that featured consistently high fractional ranks across different nations and PISA waves. Table 6.3.20 presents two pairs of questions S128Q03T/S213Q01T and S128Q03T/S270Q03T from S128 “Cloning”, S213 “Clothes” and S270 “Ozone” testlets, respectively. These two pairs were the only ones with items that were both released. An explanation of the possible positive LID between these items was offered in section 5.4.1.3. All three items were found to represent a similar format of the questions in which Yes/No judgements of the scientific merit of statements is required from students. None of the remaining items shown in Table 6.3.20 are available for public viewing. However, items S415Q07T and S415Q08T from “Solar Panels”, S466Q07T from “Forest Fires”, S495Q04T from “Radiotherapy”, and S508Q02T from “Genetically Modified Crops”, which featured in different combinations, shared the same item format type (Complex Multiple Choice) and scientific competency (Identifying scientific issues) being evaluated. Furthermore, these two item characteristics match items S128Q03T, S213Q01T and S270Q03T, suggesting that perhaps all these items are aimed to test scientific judgement and share the same Yes/No response format.

Table 6.3.20 Fractional ranks of RCs expressed as a percentage for pairs of science items that indicate between-testlet positive LID with some degree of cross-country and cross-wave consistency

PAIR_ID	INT	AUS	AUT	BEL	CAN	CHE	CZE	DEU	DNK	ESP	FIN	GBR	GRC	HUN	IRL	ISL	ITA	JPN	KOR	LUX	NOR	NZL	POL	PRT	SWE
2000_S_S128Q03T_S213Q01T	99	80	100	100	100	99	85	100	96	98	99	39	100	86	84	70	93	88	96	44	98	90	95	98	100
2003_S_S128Q03T_S213Q01T	99	100	99	100	76	100	92	96	95	88	59	87	79	98	83	17	94	92	86	99	94	99	97	100	97
2000_S_S128Q03T_S270Q03T	100	67	100	88	99	96	99	92	94	99	99	100	79	66	82	94	100	97	100	86	96	99	97	95	60
2006_S_S213Q01T_S415Q07T	99	100	86	95	100	89	100	100	43	97	87	100	50	80	24	68	96	76	57	98	100	77	78	91	68
2006_S_S213Q01T_S415Q08T	98	100	99	93	100	91	63	91	88	100	87	98	90	85	58	55	99	98	89	76	55	49	67	95	72
2006_S_S213Q01T_S495Q04T	97	100	92	45	100	95	97	91	87	79	41	94	83	83	72	84	92	52	60	99	99	67	MISS	65	92
2006_S_S415Q07T_S466Q07T	100	99	93	100	97	97	100	100	45	90	99	100	99	26	98	45	100	89	49	92	98	98	94	86	99
2009_S_S415Q07T_S466Q07T	97	70	97	99	100	85	33	97	98	98	100	100	86	96	47	99	86	77	53	7	92	15	31	51	96
2012_S_S415Q07T_S466Q07T	88	92	99	85	100	94	18	51	93	93	95	54	96	99	97	55	81	26	45	92	84	55	99	77	52
2006_S_S415Q07T_S495Q04T	100	100	99	98	100	99	60	84	82	99	82	99	97	99	62	97	98	41	100	99	99	99	MISS	86	93
2006_S_S415Q08T_S495Q04T	99	100	97	98	100	93	46	99	77	100	94	98	93	58	81	99	98	97	95	97	100	87	MISS	100	98
2006_S_S466Q07T_S495Q04T	97	99	97	100	100	100	54	99	93	99	80	100	95	83	91	92	96	98	100	83	71	89	MISS	98	99
2006_S_S466Q07T_S508Q02T	100	95	100	94	98	95	100	100	75	96	78	40	94	100	62	99	100	73	49	60	95	94	78	75	99

* **MISS** identifies an item which was not administered by some countries as reported in PISA Technical Manual (OECD, 2009b) under "National item deletions."

Table 6.3.21 also reports a pair of items from different testlets for which an indication of LID was present for international calibration and the majority of national datasets. Listed in the table and used in three waves, was the pair S428Q05 from "Bacteria in Milk" and S478Q01 from "Antibiotics", which may be positively dependent due to a common topic investigated as suggested by the testlet titles and a shared scientific content of "Knowledge of science - Living systems". It is difficult to infer plausible reasons for the positive LID for three pairs, S256Q01 from "Spoons" matching with three items from "Acid Rain", S485Q02 and S485Q03, S485Q05. However, S485 was one of the released items (while S256 was not) inquiring about the source of acid in the air, following questions about experiments with acid. Perhaps the single item S256Q01 also required similar knowledge.

Reported in Table 6.3.21 and used in three PISA studies, LID for the pair of non-released items S466Q07T and S521Q06 may be caused by specific knowledge suggested by the testlets' titles of "Forest Fires" and "Cooking Outdoors", respectively. Another pair from two testlets used in four PISA waves and featured in Table 6.3.21 is S269Q04T from "Earth's Temperature" and S326Q04T from "Milk". Both items are of "Complex Multiple Choice" item formats as well as targeting science competence of "Explaining phenomena scientifically". It is difficult to offer plausible speculations for the presence of LID for the remaining two item pairs listed in the table, S413Q05 from "Plastic Age" , S508Q03 from "Genetically Modified Crops", S495Q02T from "Radiotherapy" and S510Q01T from "Magnetic Hovertrain", as those testlets are being kept confidential and no overlapping items characteristics, apart from matching item formats, are present[51].

[51] Item characteristics information sourced mostly from (Mandíková & Bašátková, 2008; OECD, 2009b, 2014b, 2015a, 2015b, 2015c)

Table 6.3.21 Fractional ranks of RCs expressed as a percentage for pairs of science items that indicate between-testlet positive LID with some degree of cross-country and cross-wave consistency (cont.)

PAIR_ID	INT	AUS	AUT	BEL	CAN	CHE	CZE	DEU	DNK	ESP	FIN	GBR	GRC	HUN	IRL	ISL	ITA	JPN	KOR	LUX	NOR	NZL	POL	PRT	SWE
2006_S_S256Q01_S485Q02	96	96	54	84	86	90	14	95	49	74	93	98	75	48	91	97	99	99	95	93	97	100	94	89	28
2006_S_S256Q01_S485Q03	100	98	100	99	91	91	86	36	83	92	95	99	100	91	100	97	100	100	100	88	98	100	100	82	45
2006_S_S256Q01_S485Q05	98	97	90	97	88	68	88	43	86	99	96	99	98	86	99	62	81	66	82	92	98	96	87	100	95
2003_S_S269Q04T_S326Q04T	94	88	100	64	100	99	99	97	22	97	98	98	99	98	89	16	100	87	88	99	100	99	99	94	99
2006_S_S269Q04T_S326Q04T	99	70	99	100	87	68	60	95	61	100	30	88	97	97	42	100	95	97	100	83	79	50	66	90	100
2009_S_S269Q04T_S326Q04T	98	100	76	74	100	88	89	64	100	99	68	99	97	95	97	87	99	97	91	100	93	72	90	44	95
2012_S_S269Q04T_S326Q04T	100	93	98	98	99	98	85	96	99	100	98	85	97	72	47	89	97	76	99	98	98	98	96	86	59
2006_S_S413Q05_S508Q03	98	98	86	97	97	91	99	99	82	85	99	78	86	30	99	78	94	96	31	25	98	99	67	96	72
2006_S_S428Q05_S478Q01	96	98	94	100	99	100	92	98	100	93	83	99	98	100	79	98	99	86	73	98	99	98	88	93	87
2009_S_S428Q05_S478Q01	100	98	99	98	100	100	100	99	100	100	98	100	85	99	41	100	84	100	36	100	100	91	77	54	99
2012_S_S428Q05_S478Q01	100	100	99	62	100	100	70	100	100	99	99	100	93	92	94	97	100	97	91	100	100	99	67	98	100
2006_S_S466Q07T_S521Q06	100	95	100	97	99	99	100	22	64	99	99	99	91	15	98	100	72	63	65	99	70	87	48	100	41
2009_S_S466Q07T_S521Q06	97	96	97	97	100	92	72	96	98	100	28	80	97	97	97	86	100	58	72	96	76	81	97	99	72
2012_S_S466Q07T_S521Q06	98	99	94	92	99	90	98	81	98	100	87	98	95	89	58	89	100	73	39	88	61	98	91	93	77
2006_S_S495Q02T_S510Q01T	99	100	53	99	100	40	100	92	95	99	95	99	89	13	73	99	95	96	93	88	100	99	98	66	81

Negative LID suggested by close to zero fractional ranks of RCs is reported in Table 6.3.22. Explaining the likely causes of negative LID among the first six item pairs from PISA 2003 is offered below. Items S114Q03T, S304Q03A, S304Q03B, S326Q01 and S326Q01 were all shown, in previous chapters, to produce within-testlet positive LID among their items. However, in Table 6.3.22 they are now reporting negative LID between them. As discussed before, this may be due to ideas proposed by Habing and Roussos (2003) and van Rijn and Rijmen (2015). Only for the sake of comparison, the same pairs of items from PISA 2006, with science as the main investigated domain, are also reported in the centre of the Table 6.3.22. Negative LID among these five items is not present, possibly due to PISA 2006 incorporating many more items.

The table also presents four pairs of items from S114 "Greenhouse" and S195 "Semmelweis' Diary". Testlet S114 was allocated in PISA 2000 to cluster S1, while items from S195 were allocated to cluster S3. Both clusters appeared together only in Booklet 8, in which science items followed mathematical questions. Perhaps students were running out of time before the testing break and selectively allocated their efforts to deal with the lengthy testlets, as introductions and graphs were used in those items that have been released. The last listed in Table 6.3.22 PISA's 2006 item pair S425Q02 from "Penguin Island" and S426Q05 from "The Grand Canyon" also points to negative LID. Both items are allocated to the booklet in which whole two hours of testing time was dedicated to science. The item from "The Grand Canyon" was placed at the beginning of the booklet while a more difficult question about "Penguin Island" was in the last 30 minutes of testing.

Table 6.3.22 Fractional ranks of RCs expressed as a percentage for pairs of science items that show between-testlet negative LID with some degree cross-country and cross-wave consistency

PAIR	INT	AUS	AUT	BEL	CAN	CHE	CZE	DEU	DNK	ESP	FIN	GBR	GRC	HUN	IRL	ISL	ITA	JPN	KOR	LUX	NOR	NZL	POL	PRT	SWE
2003_S_S114Q03T_S304Q03A	6	1	29	9	1	8	22	7	17	3	1	13	2	0	83	1	2	9	19	6	2	2	17	11	30
2003_S_S114Q03T_S304Q03B	1	10	40	2	6	16	2	11	8	8	2	2	6	6	13	1	21	5	17	48	21	7	13	2	25
2003_S_S304Q03A_S326Q01	5	5	15	19	5	7	2	6	16	4	6	5	0	48	99	10	0	2	6	56	2	26	61	MISS	11
2003_S_S304Q03A_S326Q02	1	10	9	26	3	3	2	1	9	0	12	6	2	35	46	31	1	21	2	8	1	7	47	10	3
2003_S_S304Q03B_S326Q01	1	13	21	12	1	0	3	1	27	1	1	1	1	91	26	1	1	5	19	29	1	4	67	MISS	0
2003_S_S304Q03B_S326Q02	6	2	19	15	8	2	1	16	5	1	1	3	1	34	15	20	3	4	9	6	7	67	31	9	2
2006_S_S114Q03T_S304Q03A	37	75	86	34	73	13	14	72	83	84	73	75	37	64	86	68	8	21	39	13	49	36	51	21	55
2006_S_S114Q03T_S304Q03B	22	24	37	12	39	25	46	72	27	71	78	72	52	31	85	72	16	43	7	11	69	33	33	80	11
2006_S_S304Q03A_S326Q01	11	71	8	30	11	37	22	38	26	7	66	48	5	25	43	28	6	6	11	20	5	67	4	37	50
2006_S_S304Q03A_S326Q02	17	33	44	24	2	23	64	42	11	1	28	2	3	19	55	79	3	6	9	2	2	17	29	21	25
2006_S_S304Q03B_S326Q01	25	52	13	72	5	29	34	49	6	15	80	21	0	4	73	5	29	5	14	0	11	61	29	41	12
2006_S_S304Q03B_S326Q02	19	5	12	39	5	18	72	29	9	5	86	3	0	27	65	42	2	19	5	0	4	23	30	68	22
2000_S_S114Q03T_S195Q05T	0	5	1	1	0	5	11	1	4	21	2	3	1	0	13	9	7	1	4	7	2	54	3	1	0
2000_S_S114Q03T_S195Q06	1	21	16	38	5	8	3	11	3	0	6	1	3	22	50	1	76	3	71	5	8	15	5	3	9
2000_S_S114Q04T_S195Q05T	0	23	5	0	0	4	0	3	4	24	1	5	4	6	1	15	6	2	12	5	0	3	3	1	1
2000_S_S114Q04T_S195Q06	1	20	6	13	5	2	4	1	12	1	16	15	5	26	21	0	24	3	3	23	20	33	15	14	10
2006_S_S425Q02_S426Q05	0	1	14	11	4	0	6	0	1	0	1	4	1	11	27	35	1	7	48	0	1	2	11	6	13

* **MISS** identifies an item which was not administered by some countries as reported in PISA Technical Manual (OECD, 2005b) under "National item deletions."

As was the case in mathematics, there were a number of item pairs with fractional ranks varying between opposite extremes between countries. Table 6.3.23 shows these pairs in all the waves in which they were used. A common regularity among these items is that they are very different in difficulty with the smallest discrepancy being 51% on a scale utilising the international sample prevalence of correct responses.

Furthermore, a closer look at the booklet allocations reveals that eight pairs with more difficult items are placed in the last half an hour of testing, while one pair has both items placed in the second half of the testing time. The results in this table may be in support of selective effort or time allocation by students taking part in the PISA assessment.

Table 6.3.23 Fractional ranks of RCs expressed as a percentage for pairs of science items that show inconsistent pattern of positive or negative LID for different nations

PAIR	INT	AUS	AUT	BEL	CAN	CHE	CZE	DEU	DNK	ESP	FIN	GBR	GRC	HUN	IRL	ISL	ITA	JPN	KOR	LUX	NOR	NZL	POL	PRT	SWE
2006_S_S256Q01_S527Q01T	89	70	26	99	7	47	21	13	96	80	37	65	100	0	29	43	57	1	1	99	22	99	93	100	58
2009_S_S256Q01_S527Q01T	2	8	96	27	96	13	13	56	0	26	99	0	28	52	0	89	98	21	94	28	34	29	5	99	40
2012_S_S256Q01_S527Q01T	4	27	2	7	0	2	3	99	0	5	1	32	69	26	98	1	92	4	100	98	1	5	29	3	19
2006_S_S256Q01_S519Q01	99	43	24	12	44	75	82	77	93	65	55	82	100	97	78	100	56	0	5	48	61	3	8	31	6
2009_S_S256Q01_S519Q01	90	88	2	23	97	35	37	85	2	17	69	93	13	94	94	62	86	24	44	7	59	3	94	2	13
2012_S_S256Q01_S519Q01	13	94	7	8	13	91	1	72	96	44	76	8	38	4	83	0	59	4	91	16	47	73	5	36	24
2009_S_S131Q04D_S521Q06	7	92	0	81	0	20	50	98	62	8	14	100	9	86	18	56	4	96	6	59	98	96	96	96	6
2012_S_S131Q04D_S521Q06	90	98	2	3	0	100	0	83	25	0	95	100	7	15	38	48	18	20	100	96	57	10	44	99	56

6.3.7 Summary

The utilisation of fractional ranks of the residual correlations has its limitations as the close to 100 or 0 ranks may not relate to outlying high or low RC in all countries. However, in light of the cross-national and cross-wave inadequacy of using a single fixed cutpoint, this approach offered the possibility of looking into the consistency of cross-national positive and negative LID. As a majority of the PISA items are not available to the public, the practical explanation behind some of the observed results is limited. However, it is clear that in all three domains there were testlets written in a way that facilitated positive within-testlet dependency and which proved to be consistent in international data, all countries and in multiple waves in which items were reused. Relevant subsections also speculated on testlets which may suggest differential testlet functioning where on the one hand, most countries showed positive LID, and a few did not, or on the other hand where most countries did not show it, but a few did, in fact, reveal high positive LID. A strong case as to whether observed inconsistent cross-national positive LID prevalence is related to school curricula or teaching and learning practices, could be made only with an extensive knowledge of each participating nations' school systems and aquiring access to the items themselves. Each domain in which within-testlet LID was found also revealed testlets with no positive LID for any countries, which is a desirable testlet property from the perspective of fulfilling a local item independence assumption. The discussion about item pairs from different testlets started in each domain by highlighting the possibility of positive dependency, suggesting a specific domain knowledge, aquired skill or item format as plausible explanations. The discussion about negative LID started with tables suggesting that, as indicated preivously (Habing & Roussos, 2003; van Rijn & Rijmen, 2015), some negative LID may be a mathematical artefact of the presence of positive LID. Alternative possibilities for negative LID drivers were reported, indicating the relative location of the items during the testing time, as well as the difference in items' difficulties.

CHAPTER 7 DISCUSSION, CONCLUSIONS AND IMPLICATIONS

This chapter aims to offer a discussion of findings leading to a review of limitations. Practical implications of this research to PISA developers are offered along with the suggestions for future research. The chapter concludes with the overall conclusions.

7.1 Discussion of findings

7.1.1 Research aim 1 - Description of PISA's testlets

This research aim intended to give an efficient graphical overview of the testlets used in PISA's three cognitive domains and its five implementations. It seems that, to date, no review of cross-wave usage of mathematics, reading and science items and testlets has been published.

The results reported in Chapter 4 suggested that testlets with a larger number of items were used more frequently for reading, compared to mathematics and science. This observation is especially prominent in the initial PISA study of 2000. The descriptive graphs highlighted that very few single-item testlets were utilised to test reading and science but were frequently applied in mathematics' assessments. Another conclusion demonstrated by the high-low-close charts was that the range of item difficulty within testlets differed considerably for some testlets even when comparing the same sized testlets. This raises the question of whether the testlets should be organised with items of similar difficulty aiming so that whole testlets target a limited range of ability levels. Alternatively, the testlets can be designed so they incorporate items with a broad range of item difficulty which may increase the possibility of selective time allocation, which according to Yen (1993), produces negative LID. Finally, describing the PISA testlets offered succinct one-stop visualisations of the patterns of testlets used for cross-wave linking, highlighting that, for example, only eight testlets were used in the reading assessment of PISA 2003 and 2006.

7.1.2 Research aim 2 - LID in data from PISA's international calibrations

There were four research questions grouped under research aim 2 that utilised data from PISA's international calibrations, and the results pertaining to these questions are presented in Chapter 5.

Firstly, the prevalence of positive and negative LID, according to the non-IRT LID index used in this study, is provided. A very limited number of publications (reviewed in detail in section 2.9) indicated the presence of LID in PISA, and in doing so the various authors used only small subsets of the PISA data. However, this research revealed rates of positive and negative LID which were

subject to meta-analysis while aggregating across five PISA waves. The prevalence of LID was found to be just as high for mathematics as it was in reading, while it was lower for science. The fact that certain cognitive domains are targeted in the specific waves of PISA assessments does not appear to be associated with inflated or deflated LID prevalence. Outlying large negative residual correlations were about 2.5 times more prevalent than positive outlying RCs. As no other research has undertaken such a comprehensive approach to estimating LID prevalence in PISA, this part of the research offers a novel contribution to the literature yet limits possibilities for cross-validation with existing publications.

Secondly, LID prevalence was investigated in relation to the within-testlet or between-testlet location of item pairs. Among the item-pairs belonging to non-singular mathematical testlets, the within-testlet positive LID prevalence for mathematics was 43% (28%, 59%) with 27 out of 39 non-singular mathematics testlets having at least one pair of its items for which the RC exceeded +0.1. While within-testlet LID in reading is expected and attributed in the literature to common introductory reading passages, the results re-iterated above, regarding the within-testlet dependency in mathematics, are novel. Further, new results are offered by reporting the positive dependency among items from different testlets.

Thirdly, the plausible drivers of the LID are explored in two ways by an in-depth qualitative review of the released items for which LID was identified, and through multilevel logistic regressions predicting positive and negative dependency utilising a range of predictors quantifying item pair characteristics.

The qualitative investigations of within-testlet LID in mathematics suggested that common testlet stimuli are not likely to be the only reasons for positive LID with some items showing item chaining or common mathematical abilities as more LID indicative, compared to shared testlet introductions. The between-testlet positive dependency qualitative investigations, made possible through access to the publically released items, also gave some plausible explanations to some high residual correlations, pointing to the very specific skill of, for example, mean calculation or geometrical inference from the dice as a likely LID cause. Positive within-testlet LID in reading was primarily related to the common introductory text as far as it could be determined in released items. The between-testlet positive LID was more difficult to explain, due to the very limited number of item pairs that were readily available for review. This was also the case for the science domain for which items are more closely protected and even fewer items are released to the public. The existence of positive within-testlet dependency in science was, it is argued, due to shared introductory figures for multiple items within some testlets. The investigations of the between-

testlet positive dependency in science proved to be very interesting, suggesting that the Yes/No style of complex multiple choice item format, which aimed to evaluate the science competency of "Identifying scientific issues", could drive between-testlets positive LID. This insight into the science domain was particularly important given the scarcity of other published research about LID in science related PISA items. The results of the qualitative investigations largely concurred with other publications discussing LID in PISA's mathematics or reading items.

Qualitative investigations of negative dependency proved to be more difficult and more speculative. Negative LID in the between-testlets pairing of items appeared to involve some of the items for which one was quicker for the student to investigate while the other involved much more cognitive investment. As a consequence, for some of the qualitative investigation of negative LID, it can be argued that selective time and effort allocation may be at least partly responsible for negative LID as per Yen's (1993) suggestions. On the other hand, it was also visible that between-testlet negative dependency very frequently involved items which also had a high within-testlet positive dependency. This would be more in agreeance with Habing and Roussos (2003) conclusions showing the mathematical need for negative LID when positive dependence is present. The qualitative investigation of LID dependency was limited due to the lack of access to the majority of items. However, one of the contributions of this research, with regard to this particular research question, is that the results for all items are reported in electronic appendices, facilitating the possibility of further subsequent in-depth reviews by researchers with full access to PISA's confidentialised cognitive questions.

The quantitative part of addressing this research question for positive LID confirmed some of the observations from qualitative investigations. This was particulary apparent for mathematics predictors of LID, namely item pair location or a common mathematics strand such as geometry. The quantitative results for mathematics also produced an interesting finding, suggesting that the odds of positive LID are high for a pair of items that differ considerably in difficulty in comparison to the reference of moderate difficulty item pair. It is speculated that external assistance or academic cheating could be at play in this case, as this was suggested by Yen (1993) as one of LID drivers, although in this study using secondary data it is not possible to test this speculation. The conclusions of the logistic regression models pointed to various possibilities of selective time and effort allocations, predicting at least partly negative dependency. The quantitative part of addressing this research question also offers a possibility that investigating the causes of LID may offer us an insight into the strategies students are using while exposed to testing.

The fourth and final research question aimed to use international calibration data to examine cross-

wave LID consistency. The PISA cross-wave LID consistency aspect of this research, to my knowledge, has not been previously investigated. As demonstrated in section 2.6, local item dependency can indeed have a negative effect on cross-wave equating procedures. Section 5.4.1 discusses in detail that in all three cognitive domains there is within-testlet dependency that has been shown to be consistent across all the PISA waves which used these testlets. Some of the investigated item pairs produced LID in four or five PISA waves that used them. Furthermore, the importance of the cross-wave LID consistency could be further elevated depending on how many linking testlets are impacted. For example, after reading was the main targeted cognitive domain in PISA 2000, eight testlets were retained for PISA 2003 with five of them having at least one item pair featuring positive within-testlet LID in both waves.

7.1.3 Research aim 3 - LID in data from PISA's national calibrations

This research aim has four supplementary research questions. The first two questions relate to comparing the levels of LID prevalence across 24 OECD countries and proposing arguments as to why some countries may display higher levels of local item dependency. It was found that in regard to between-testlet positive LID and looking across all three cognitive domains that the countries preforming highly in PISA, such as Finland, Japan, Korea and in the case of reading Ireland, featured in some PISA waves as displaying higher levels of this type of LID. It was suggested that perhaps students from these countries were taught higher order comprehension skills. Hence they can take advantage of shared content or formats of items even if they are from different testlets. It could be speculated that perhaps the educational evaluation systems in these countries are more aligned with PISA type tests which would emerge as a between-testlet dependency. In regard to positive within-testlet dependency, Greece frequently featured as an outlier, particularly in reading and science. In regard to negative LID, Greece appeared again displaying levels higher than other countries of this type of LID. It is possible that this was an artefact of higher levels of positive LID being observed in Greece, in line with the justifications made by Habing and Roussos (2003). On the other hand, as students from Greece were poorer performers in the PISA study compared to other investigated countries, perhaps they were implementing some of the selective time and effort allocation strategies more frequently and that could lead to negative LID (Yen, 1993), resorting to only attempting questions that appeared to be relatively easy. However, it needs to be acknowledged that more defensible arguments explaining why some countries present higher LID levels could only be put forward by researchers who possess an in-depth understanding of these particular nations' educational systems, challenges and curriculums.

The second half of Chapter 6 addressed research questions about cross-national consistency in LID as well as the possibility of differential testlet functioning. The results confirmed high levels of

cross-national positive LID consistency for selected item pairs originating within and between testlets. At the same time, there were indications for differential testlet functioning by a qualitative review of item pairs which either did not show positive LID only for some countries or featured LID in a limited number of countries. Testlets which were consistently non-presenting LID in national or international databases were also reported, as they may be worth reviewing in depth by the PISA test developers. Cross-national consistency in regard to negative dependency was less clear, and its drivers were argued to be the effect of positive within-testlet LID (Habing & Roussos, 2003).

7.2 Limitations

There is a number of potential limitations of this study which need to be acknowledged.

7.2.1 Limitations related to research aim 2

Fixed value of residual correlation as an indicator of LID

This research followed Kline (2016) recommendations in regards to the value of residual correlation indicative of local item dependency. The limitations of utilising fixed cut-points are acknowledged in regard to SEM fit indices (Heene et al., 2012) or Rasch model fit statistics (Wu & Adams, 2013). The very recent publication by Christensen et al. (2017) suggests the use of a simulation approach to identifying cut points for Q_3 dependency index and points to the cut-point dependency on sample size. This was also observed in this study in Section 6.2 rendering changes to the methodological approach while addressing research questions from research aim 3. This limitation emerged as a consequence of the recent research developments and therefore could not be incorporated during this part time PhD candidature. Also mimicking Christensen et al. (2017), the simulation approach to locating cut-points would be very challenging, particularly for countries which, in the PISA study, used very large sample sizes in the main targeted cognitive domains and used close to 100 items.

Correction due to possible negative bias in residual correlations

Some authors who investigated LID using the IRT models index suggested that mean value of residuals (Marais, 2013) or term -1/(Number of items-1) (Chou & Wang, 2010) should be added to the IRT residuals. Neither Marais nor Chou and Wang discussed sample sizes or a number of items even close to PISA's conditions. Furthermore, no publication suggesting the similar correction for residual correlations from CFA could be located. It was decided not to apply any corrections. The median value of residual correlations for international data CFAs was -0.011 while Chou's term

median value was -0.023. This would render the impact of the correction factor small for the majority of the results. However, it is acknowledged that if future research gives additional arguments for the implementation of one of these correction factors, this would indicate that this study has a slightly inflated prevalence of negative LID and a slightly underestimated positive dependency. This was to some extent indicated in this research during qualitative investigations related to the science cognitive domain when access to released items allowed the identification of item pairs showing interpretable dependency while only marginally missing the cut-points. The same limitation applies to analyses addressing research aim number 3.

Treating non-reached cognitive items as missing

The issue of treating non-reached items as missing and therefore underestimating the dependency as per Monseur et al. (2011) conclusions was mentioned in section 5.2. The same limitation applies to the analyses addressing research aim number 3.

Multilevel logistics regressions

There are four possible limitations in regard to multilevel logistic regressions.

Firstly, as mentioned in section 3.3.1, the interaction terms were not investigated in the multilevel logistic regressions reported in section 5.4.2. This was due to a small number of events in some of the logistic regressions raising the possibilities of computational challenges while estimating the interactions.

Secondly, with regard to explaining negative dependency during the modelling process, a new possible predictor was identified, backed by published research (Debeer & Janssen, 2013). While this new predictor quantifying the relative location of item pairs through the testing time could be prepared from available information, its preparation wouldn't be trivial in the timeline available and given the number of PISA waves and cognitive domains involved.

Thirdly, cognitive domain-specific item properties required extensive categorising and variable management prior to being included in the multilevel logistic regressions. The aggregations were mostly driven by small counts and a large number of categories. For example, in reading, 15 different languages were listed in technical manuals in relation to original item submissions to the PISA consortium. From the perspective of investigating pairs of these items, if the original languages were to have been retained, a variable reflecting source language effects on the presence of LID would have been required to model a variable with 120 levels. Consequently, the languages were grouped into their language families such as Germanic Family or Hellenic/Italic family based

on Wikipedia's classification resulting in the final language related independent variable having10 levels. Changes in item format classification across the five PISA waves provided another challenge. The item format classification was changed twice throughout the PISA study with a major shift driven by the introduction of computer-based assessment. Item format was recorded into fewer categories dictated by the need to limit the number of parameters in the model. "Simple Multiple Choice", and "Complex Multiple Choice" item formats have been used consistently across all five waves and remained unchanged. Items that were coded in different PISA waves as "Short Response", "Closed Constructed Response", "Constructed Response Auto-coded", "Constructed Response Manual" were re-labelled as "Short response" as the literature indicates that this is a key common denominator for all these item types (Neidorf, Binkley, Gattis, & Nohara, 2006; Stacey & Turner, 2015a). Similarly, "Constructed Response Expert" and "Open Constructed Response" were combined together as, in essence, they required from students extended written responses. This example shows the challenges not only related to reducing the number of categories but also to maintaining comparable item properties across five waves. All the variables modifications are elaborated in section 5.4.2 and are recorded in electronic IBM SPSS syntax files, which are available to view upon request. The data management, mentioned above, was meticulous and should not induce bias in the results of the logistic regression models.

Fourthly, as is the case in any non-experimental model-based research, the selection of independent variables for the models is non-exhaustive. Despite the extensive searches reported in section 3.3.2.4 "Search for information about cognitive items" there may be other characteristics of item pairs which could serve as predictors of LID yet were not located and quantified for the multilevel logistic regression models.

Conservative choice of dual-index LID for section 5.4.1

As mentioned in section 5.4.1, due to a large number of residual correlations that would be required to investigate qualitatively, a second LID index was used for this part of the research to focus the investigation and make the qualitative examinations feasible. As it can be seen in Figure 5.4.1, there were item pairs with high RCs which did not have high modification indices. These item pairs were not checked in regard to belonging to the groups of released to the public items, yet some could be eligible for in depth qualitative review.

Non-inclusion of the PISA 2015 data

At the time when the bulk of analyses in this part-time PhD were undertaken the cognitive data from PISA 2015 were yet released. It was not viable to include PISA 2015 data at the time of their

release to the public. The same limitation applies to analyses addressing research aim number 3.

7.2.2 Limitations related to research aim 3

Zero frequency cells estimation warnings in Mplus

Final data including all item pairs for three cognitive domains, five PISA waves and 24 investigated countries consisted of just over 727000 residual correlations as extracted from CFAs conducted in Mplus software. A very small proportion of these (0.33%, n=2418) produced a warning message in the Mplus output. There were two types of warnings:

Warning Type 1 – "WARNING: THE BIVARIATE TABLE OF [ITEM A] AND [ITEM B] HAS AN EMPTY CELL."

Warning Type 2 – "WARNING: THE SAMPLE CORRELATION OF [ITEM A] AND [ITEM B] IS [ESTIMATE VERY CLOSE TO 1 OR -1] DUE TO ONE OR MORE ZERO CELLS IN THEIR BIVARIATE TABLE. INFORMATION FROM THESE VARIABLES CAN BE USED TO CREATE ONE NEW VARIABLE."

Both warnings relate to the presence of zero frequency cells in the bivariate contingency tables involving pairs of items. The reasons reported in the literature concerning the presence of zero cells may relate to the small sample size as well as the existence of extreme thresholds (Savalei, 2011), indicating that at least one item involved in the estimation of bivariate correlation was either extremely easy or extremely difficult for students.

The number of warnings reported for each combination of wave and domains varied. The electronic appendix (Electronic Appendix 7.2.1_List of all CFA warnings produced from national calibration data.xlsx) reports all the warnings. Out of the total of 2418 pairs of items for which Mplus issued a warning 2205 (91%) were of Type 1 informing about the presence of an empty cell.

There appears to be a limited literature in regard to how to handle zero frequency cells for estimations involving categorical data and WLSMV estimators. Suggestions (non-supported by any citations) from the Mplus discussion board are somewhat inconsistent, pointing to either the removal of one item or the removal of both items contributing to the generation of zero frequency warning (Muthén & Muthén, 2007-2017). This suggestion is contrary to the paper by Savalei (2011) that investigated the issue, deciding which of the two typically used methods of dealing with the problem should be used, i.e. (NONE – ignoring the zero frequency warnings versus ADD – adding small values to the zero frequency cells). The paper concludes that for categorical items with three or more categories no special treatment is required. In the case of binary items, the author

gives a slight preference to the ADD method, which the default approach (Muthén & Muthén, 1998-2015, p. 831) in the Mplus estimation used for CFAs in this chapter (i.e. to add 0.5 divided by the sample size to the zero frequency cells). The very conservative approach suggested on the Mplus discussion board is not plausible from the perspective of cross country and cross PISA waves comparisons. It would bring about considerably reduced models with a varying number and selection of items from country to country. Given the previously reported very small proportion of items' pairs resulting in the warnings generation, the Mplus default approach to zero cell problem was retained.

There were also seventeen Type 1 CFAs warnings for the international data analyses, constituting 0.056% of all residual correlations used in Chapter 5 investigations. The electronic appendix (Electronic Appendix 7.2.2_List of all CFA warnings produced from international calibration data.xlsx) reports all of these.

Covariance coverage problem when dealing with missing data by design

As stated in Mplus manual

> The output for this analysis [missing data investigation] produces the number of missing data patterns and the proportion of non-missing data, or coverage, for variables and pairs of variables. A default of .10 is used as the minimum coverage proportion for a model to be estimated. This minimum value can be changed by using the COVERAGE option of the ANALYSIS command. (Muthén & Muthén, 1998-2015, p. 490).

Because of the booklet design, only subsets of students were exposed to item pairs producing the "missing by design" data patterns. For datasets from cognitive domains which were targeted in PISA waves some item pairs reported a covariance coverage value below the Mplus default and therefore the default was changed to 0.05.

Outliers based on selected 24 OECD countries

Identification of countries which present outlyingly high levels of LID, as reported in section 6.2, was made using a relatively small group of 24 OECD countries, which raises issues related to the statistical power of the test for identifying the outliers and the reproducibility of results should other countries participating in PISA also be included on a later occasion.

Slightly different sets of cognitive items used in national datasets

In the case of national calibrations, the countries were allowed to opt out of using some cognitive items, and these exclusions are reported in all five PISA Technical Manuals under the

"National Item Deletions" headings. Items listed as not used at all for the particular country were excluded from the corresponding national CFAs. However, fifteen CFAs from the initial round of the 390 Mplus estimations still did not converge. The section directly below reports all non-convergence situations along with the comment as to how the issue has been addressed.

PISA 2000 Mathematics - FRANCE

For example in PISA 2000 when mathematics was investigated the results for France were not obtained. The plausible reasons for this were investigated. Firstly it was checked whether some items not listed in the technical manual were additionally excluded in France's original cognitive dataset, and this was found not to be the case. However, cross tabulations for items showed that some pairs of items (for example M150Q01 and M034Q01T) produced no counts indicating that in France not a single child was exposed to these two items at the same time. This cross tabulation was an exception, compared to all the other countries. Electronic appendix (Electronic Appendix 7.2.3_Selected cross tabulations for mathematical data from France in PISA 2000.xlsx) shows a cross-tabulation of a few more mathematical cognitive items also presenting this anomaly. The technical manual for PISA 2000 does not give a plausible explanation, but acknowledges that "Results from the inter-country reliability study indicated an unexpectedly high degree of variation in national ratings of open-ended items." (Adams & Wu, 2002, p. 185) One possibility for this peculiarity in French data may be due to a different arrangement of item allocation to the booklets that was used. Another reason may be that at the early data entry stages the labelling of items may have been confused, so that the responses to M150Q01 or M034Q01T do not correspond to the raw data collected at schools. Therefore France was removed from the cross-national comparisons altogether.

PISA 2000 Reading - ITALY

PISA 2000 Technical Manual (Adams & Wu, 2002, p. 185) states that the R219Q01T item was chosen to be deleted by the Italian PISA implementation team. However, it appears that R219Q01E item has only missing data while R219Q01T was responded to by the students. The CFA analyses were re-run again with a correction of the input files to exclude R219Q01E in favour of R219Q01T

PISA 2000 Science - NORWAY

PISA 2000 Technical Manual (Adams & Wu, 2002, p. 185) lists Iceland and Netherlands as countries that opted out of using the S268Q02T item. After the initial CFA estimation produced an error "Categorical variable S268Q02T contains less than 2 categories" it was confirmed in the raw data that all cases are listed missing for this item, with this also being the case for Norway. The

analysis was re-run without the item.

PISA 2003 Mathematics - ICELAND

While the PISA 2003 Technical Manual (OECD, 2005b, p. 190) mentions that item M144Q03 was deleted only from Booklet number 4, this caused a number of errors in CFA estimations related to zero covariance coverage. The item was removed and CFA re-estimated.

PISA 2003 Mathematics - DENMARK

The reason for non-convergence was the same as for Iceland, mentioned previously, with the exception that the item M273Q01 was deleted from one of the booklets.

PISA 2003 Mathematics - USA

Item M505Q01 was found to be quite difficult for the American PISA participants and the cross-tabulation with another question M413Q01 resulted in three out of four cells with zero counts. This, in turn, caused computational problems in estimating the correlation between these two items. To address this issue M505Q01 was removed from the CFA re-run.

PISA 2006 Mathematics - ICELAND

The technical manual for PISA 2006 (OECD, 2009b, p. 216) reported that item M442Q02 was excluded in Iceland from one of the booklets. The initial CFA run incorporated this item, but its removal from Booklet number 7 proved to generate zero covariance coverage errors. CFA was re-run without this item.

PISA 2006 Reading - the USA

Due to an error in printing the booklets in the USA, the PISA consortium decided to exclude the American reading data from the cognitive database (OECD, 2007a). No CFA results are available for this country. The USA was removed from cross-national comparisons altogether.

PISA 2009 Mathematics - POLAND

Although the PISA 2009 Technical Manual (OECD, 2012, p. 196) reports that item M442Q02 was excluded from one booklet, the initial run of CFAs allow for this item. This item produced a lack of covariance coverage with few other items and caused a failure to estimate this analysis. This item has been removed, and the CFAs were re-run.

PISA 2009 Reading - HUNGARY ICELAND JAPAN IRELAND PORTUGAL

Once again the runs of CFAs for all these countries, with items listed in the PISA 2009 Technical Manual (OECD, 2012, p. 196-7) as partially excluded for one booklet, did not converge. The excluded items listed in the manual were removed, and the Mplus analyses were re-estimated.

PISA 2012 Mathematics - ICELAND

Item M995Q03 has not been listed in the National Deletions section of the Technical Manual (OECD, 2014b) as having been removed from any booklets. However, it caused a considerable number of lack of covariance coverage error messages similar to other countries when items were selectively excluded from some booklets. This item was removed from the CFA input file and re-run.

PISA 2012 Science – FRANCE

Two items S131Q02D and S131Q04D were excluded from Booklet 1 for French students yet retained in the initial CFA that caused non-convergence. Both questions were excluded in the second run of CFA.

Using fractional ranks of residual correlations

In the second half of Chapter 6, it was necessary for the sake of cross-national consistency to use fractional ranks as opposed to residual correlations. While it is assumed that this part of the research offers valuable insights into cross-national LID consistency, the direct comparability to results from international calibration datasets cannot be ascertained.

7.3 Practical implications for PISA developers

There are immediate practical implications that PISA developers could consider, based on the results of this research aimed at the reduction of LID as suggested by Mazzeo and von Davier (2008).

A number of suggestions regarding how LID can be avoided can be offered. Firstly, the thesis identifies LID presence in some testlets, but it also locates the testlets which, despite using common prompts, do not reveal a violation of this IRT assumption. As the majority of the items and their corresponding testlets are not released to the public at this time, only researchers working on PISA studies and who have access to the non-released items will be able to use the findings reported in this thesis and locate non-LID testlets which could be reused for cross-wave linking while avoiding

the LID indicating items shown in this research. Secondly, the PISA test development team with unrestricted access to items can take advantage of the electronic appendices which accompany section 5.4.1 and seek to extend the qualitative investigations of LID drivers to all item pairs including non-released questions. Thirdly, more consideration by the PISA test development teams can be given to the possible causes of item dependency that are not related to common stimuli for items not located in the same testlet. The results of this study reported in section 5.4.2 suggest that, for example, combinations of items of the open-ended response type are more likely to present negative LID while item pairs which differ considerably in their difficulty tend to show positive LID. Further, positive LID was more likely found due to the involvement of a common mathematical formula or the requirement to identify the scientific nature of statements. These non-testlet related LID drivers could be considered in the selection of items for future PISA studies. Fourthly, consideration could be given to using Testlet Item Response Theory models (Wainer et al., 2007) in scaling of the cognitive PISA data.

A procedure in which the quality of national and international calibration data is investigated could also potentially incorporate LID assessments. The LID investigation could be included in the early stages when field testing is undertaken with the initial pool of items. This may be particularly important, as the naming of items used in PISA studies such as R404Q10A, R119Q09T, or S415Q08T suggests that as many as ten, nine or eight, respectively cognitive questions were considered and possibly tested, yet not included in the final study.

7.4 Suggestions for future research

This study focused on one LID index namely residual correlation from factor analysis. However, other indices can also be used for LID detection. Throughout the duration of this part-time PhD candidature the Q_3 index gained popularity. There is only one other publication (Monseur et al., 2011) explicitly targeting the investigation of LID in PISA and the authors used Q_3. While the results of this PhD investigation largely concur (see Section 5.4.1.2 and 5.4.1.4) with the findings of Monseur et al. (2011), these two research projects have a very limited overlapping data range. Subsequent research could focus on using the Q_3 index with a more extensive cross-wave and cross-domain investigation of LID in PISA. Recent methodological advances (Christensen et al., 2017), related to critical values of Q_3 for LID detection values, could be extended to include critical values investigations for studies of the magnitude of PISA's with its large sample sizes and the large number of cognitive items used. These new critical values of Q_3 could, in turn, be used in quantifying the prevalence of LID in PISA.

This research not only focused on the first five waves of PISA and three cognitive domains but

also concentrated on data from pencil-and-paper assessments and a selection of mostly developed OECD countries. There is the possibility of extending this study to the latest, recently released, data from PISA 2015 or any subsequent PISA implementations. Also, data from computer-based assessments could be investigated. Various implementations of the PISA study also examine problem-solving skills or financial literacy. The existence of LID in these other literacies could be studied. Furthermore, with the introduction of the easy booklets option from PISA 2009 onwards, LID prevalence could be compared between a standard set of booklets and the easy booklets.

There is a considerable body of literature which was reviewed in Chapter 2 pointing to various negative aspects of local item dependency. In this study, the prevalence and likely causes of LID have been investigated. What remains unknown, with the exception of limited findings from Monseur et al. (2011) and Kreiner and Christensen (2014), is the effect of LID on the country estimates or country PISA ranks. Such investigation would require an extensive set of simulation studies. Such studies could use the findings of the current research on the prevalence of and possible causes of LID in building models to be tested in the simulations.

Given that, as shown in this study, some testlets consistently reveal LID in multiple iterations of PISA, the effect of LID on the robustness of cross-wave equating in PISA could be another avenue of research. The publication by Kasper, Ünlü, and Gschrey (2014) investigating sensitivity used in the PISA IRT model to different missing data approaches, is a good example of this type of research.

The interpretation of the negative LID proved to be challenging but suggested a possibility of two different mechanisms for its creation. On the one hand, negative LID could be a mathematical consequence of positive within-testlet LID. On the other hand, the results of this study suggested that selective time and effort allocation may also be at play. With PISA 2018 transitioning to being largely computer based, consideration could be given to including technical capabilities of collecting response time data. This information could allow the investigation of patterns of time allocation to different types of questions as possible predictors of negative LID. Research utilising response time data is emerging (Bolsinova & Tijmstra, 2016; Bolsinova, Tijmstra, et al., 2017). Similarly, the effect of item placement within the testing time and its relation to negative LID could be researched.

The results of this research showed that use of pairwise LID indices might give an incomplete picture of within-testlet item dependency. Some testlets reported residual correlations for within-testlet item pairs only just exceeding the utilised cut-point, yet if all items in the testlet were flagging such marginal LID, the overall testlet dependency might not be marginal. Monseur et al.

(2011) acknowledged this issue by reporting a median within-testlet LID index. Future research could focus on developing indices that capture more robustly within-testlet dependency.

In section 6.3 a cross national perspective on LID was investigated, and differential testlet functioning was examined from a qualitative perspective, focusing on selected testlets. Future research could utilise some of the outlier detection techniques optimised for big data and multivariate applications (Finch, 2012; Nag, Mitra, & Mitra, 2005) or other approaches (Fukuhara & Paek, 2016).

The usefulness of LID indices for the purpose of detecting cheating has been investigated experimentally by Zimmermann et al. (2016). Should this study be replicated with a larger number of countries, it would be of interest to determine whether the prevalence of positive LID correlates with some of the national indices of corruption or school dishonesty.

7.5 Overall Conclusions

This thesis set out to investigate the existence of local item dependency in the PISA study aiming to provide comprehensive and generalised LID investigations by including data from the first five waves of PISA tests and three cognitive domains, namely mathematics, reading and science. The existence of LID was also examined by using PISA's international and national calibrations datasets. It is argued that this research provided evidence that the local item independence assumption is violated in PISA and the generalisability of this statement was revealed by showing cross-wave and cross national consistency in LID presence. Particularly novel are the results indicating LID in the sets of items used in cross-wave linking but also highlighting its prevalence in the mathematics domain. Also original to this research were the investigations of plausible drivers of dependency, not only undertaken qualitatively but also through statistical models predicting the LID. While the existence of LID due to common stimuli for items located within testlets was confirmed, other likely causes of dependency were suggested. Also specific to this research was an investigation of the different types of LID, i.e. positive and negative. The presence of testlets indicative of differential testlet functioning was stipulated along with identifying countries in which overall higher levels of dependency are present.

Potentially in the future, this thesis could be consulted by the many researchers involved in PISA's implementation. They will be able to re-use those items that do not reveal LID in future PISA implementations. The reporting of LID in non-released items will also offer the PISA team, through their full access to the questions, an opportunity to conduct a qualitative review of LID producing pairs of items similar to the investigations of released items conducted in this research. It

is suggested that LID investigations should be part of the PISA test development and quality testing procedures in line with the Mazzeo and von Davier (2008) recommendations. Furthermore, LID investigation may give us many insights into students' effort allocations and trains of thought while responding to the cognitive questions. While research specifically investigating the impact of LID on PISA's country rankings is scarce, two publications by Monseur et al. (2011) and Kreiner and Christensen (2014) offer initial evidence that violating this assumption can influence PISA's country ranks i.e. PISA results which are given the most attention by media. This thesis, by locating LID in five PISA waves and three cognitive domains, gives an additional argument for further research to understand LID impact on PISA's results.

REFERENCES

Abrahantes, J. C., & Aerts, M. (2012). A solution to separation for clustered binary data. *Statistical Modelling, 12*(1), 3-27.

Ackerman, T. A. (1987). The Robustness of LOGIST and BILOG IRT Estimation Programs to Violations of Local Independence. *ACT Research Report Series, 87*(14).

Adams, R. J. (2003). Response to 'Cautions on OECD's Recent Educational Survey (PISA)'. *Oxford Review of Education, 29*(3), 377-389.

Adams, R. J. (2011). Comments on Kreiner 2011: Is the foundation under PISA solid? A critical look at the scaling model underlying international comparisons of student attainment. Retrieved from https://www.oecd.org/pisa/47681954.pdf

Adams, R. J., Wilson, M., & Wang, W.-C. (1997). The Multidimensional Random Coefficients Multinomial Logit Model. *Applied Psychological Measurement, 21*(1), 1-23.

Adams, R. J., & Wu, M. L. (2007). The Mixed-Coefficients Multinomial Logit Model: A Generalized Form of the Rasch Model. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and Mixture Distribution Rasch Models: Extensions and Applications* (pp. 57-75). New York: Springer.

Adams, R. J., & Wu, M. L. (2009). The construction and implementation of user-defined fit tests for use with marginal maximum likelihood estimation and generalized item response models. *Journal of Applied Measurement, 10*(4), 355-370.

Adams, R. J., & Wu, M. L. (Eds.). (2002). *Programme for International Student Assessment (PISA): PISA 2000 Technical Report*. Paris: OECD Publishing.

Adams, R. J., Wu, M. L., & Carstensen, C. H. (2007). Application of Multivariate Rasch Models in International Large-Scale Educational Assessments. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and Mixture Distribution Rasch Models: Extensions and Applications* (pp. 271-280). New York: Springer.

AERA, APA, & NCME. (1974). *Standards for educational and psychological tests and manuals*. Washington, DC: American Psychological Association.

AERA, APA, & NCME. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York: Wiley.

Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-Practice Recommendations for Defining, Identifying, and Handling Outliers. *Organizational Research Methods, 16*(2), 270-301.

Akkari, A., & Lauwerier, T. (2015). The education policies of international organizations: Specific differences and convergences. *Prospects, 45*(1), 141-157.

Allen, S., & Sudweeks, R. R. (2001, April 10-14). *Identifying and Managing Local Item Dependence in Context-Dependent Item Sets.* Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Allison, P. D. (2008). *Convergence failures in logistic regression*. Paper presented at the SAS Global Forum, San Antonio, Texas. http://www2.sas.com/proceedings/forum2008/360-2008.pdf

Almond, R. G., Mulder, J., Hemat, L. A., & Yan, D. (2009). Bayesian network models for local dependence among observable outcome variables. *Journal of Educational and Behavioral Statistics, 34*(4), 491-521.

Amtmann, D., Cook, K. F., Jensen, M. P., Chen, W.-H., Choi, S., Revicki, D., . . . Lai, J.-S. (2010). Development of a PROMIS item bank to measure pain interference. *Pain, 150*(1), 173-182.

Anastasi, A. (1961). *Multiple choice items for psychological testing* (2nd ed.). New York: Macmillan.

Anatchkova, M. D., Barysauskas, C. M., Kinney, R. L., Kiefe, C. I., Ash, A. S., Lombardini, L., & Allison, J. J. (2014). Psychometric Evaluation of the Care Transition Measure in TRACE-CORE: Do We Need a Better Measure? *Journal of the American Heart Association, 3*(3), 1-11.

Anderson, J. O., Lin, H. S., Treagust, D. F., Ross, S. P., & Yore, L. D. (2007). Using large-scale assessment datasets for research in science and mathematics education: Programme for International Student Assessment (PISA). *International Journal of Science and Mathematics Education, 5*(4), 591-614.

Andrich, D. (1978). A Rating Formulation for Ordered Response Categories. *Psychometrika, 43*(4), 561-573.

Andrich, D. (2016). Accounting for Local Dependence with the Rasch Model: The Paradox of Information Increase. *Journal of Applied Measurement, 17*(3), 262-282.

Andrich, D., Humphry, S. M., & Marais, I. (2012). Quantifying Local, Response Dependence Between Two Polytomous Items Using the Rasch Model. *Applied Psychological Measurement, 36*(4), 309-324.

Andrich, D., & Kreiner, S. (2010). Quantifying response dependence between two dichotomous items using the Rasch model. *Applied Psychological Measurement, 34*(3), 181-192.

Arffman, I. (2010). Equivalence of translations in international reading literacy studies. *Scandinavian Journal of Educational Research, 54*(1), 37-59.

Arzarello, F., Garuti, R., & Ricci, R. (2015). The impact of PISA studies on the Italian National Assessment System. In K. Stacey & R. Turner (Eds.), *Assessing mathematical literacy: The PISA experience.* (pp. 249-260): Cham: Springer.

Asil, M., & Brown, G. T. L. (2016). Comparing OECD PISA Reading in English to Other Languages: Identifying Potential Sources of Non-Invariance. *International Journal of Testing, 16*(1), 71-93.

Baird, J., Isaacs, T., Johnson, S., Stobart, G., Yu, G., Sprague, T., & Daugherty, R. (2011). *Policy effects of PISA*. Oxford University Centre for Educational Assessment. Retrieved from http://oucea.education.ox.ac.uk/wordpress/wp-content/uploads/2011/10/Policy-Effects-of-PISA-OUCEA.pdf

Balazs, K., & De Boeck, P. (2006). *Detecting Local Item Dependence Stemming from Minor Dimensions. IAP Technical Report 0684*: IAP Statistics Network. Retrieved from https://sites.uclouvain.be/IAP-Stat-Phase-V-VI/PhaseV/publications_2006/TR/TR0684.pdf

Bao, H., Gotwals, A. W., & Mislevy, R. (2006). *Assessing Local Item Dependence in Building Explanation Tasks (DRAFT). PADI Technical Report 14*. SRI International Center for Technology in Learning. Retrieved from https://padi.sri.com/downloads/TR14_LocalDepend.pdf

Barendse, M. T., Oort, F. J., & Timmerman, M. E. (2015). Using Exploratory Factor Analysis to Determine the Dimensionality of Discrete Responses. *Structural Equation Modeling, 22*(1), 87-101.

Bartolucci, F. (2007). A class of multidimensional IRT models for testing unidimensionality and clustering items. *Psychometrika, 72*(2), 141-157.

Bautier, E., & Rayou, P. (2007). What PISA really evaluates: literacy or students' universes of reference? *Journal of Educational Change, 8*(4), 359-364.

Beaton, A. E. (1987). *Implementing the New Design: The NAEP 1983-84 Technical Report. (Report No. 15-TR- 20)*. Princeton, NJ: National Assessment of Educational Progress, Educational Testing Service. Retrieved from http://files.eric.ed.gov/fulltext/ED288887.pdf

Bennett, R. E. (2010). Technology for Large-Scale Assessment. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International Encyclopedia of Education (Third Edition)* (pp. 48-55). Oxford: Elsevier.

Berezner, A., & Adams, R. J. (2017). Why Large-Scale Assessments Use Scaling and Item Response Theory. In P. Lietz, J. Cresswell, K. Rust, & R. J. Adams (Eds.), *Implementation of Large-Scale Education Assessments* (pp. 323-356). Hoboken, NJ: John Wiley & Sons.

Berezner, A., & Timms, M. (2016, 21 March). [Personal communication on some inconsistencies and lack of clarity in the PISA OECD Technical Manuals in regard to the creation of the International calibration samples for PISA 2009 and 2012].
Berliner, D. C. (2015). The Many Facets of PISA. *Teachers College Record, 117*(1).
Best, M., Knight, P., Lietz, P., Lockwood, C., Nugroho, D., & Tobin, M. (2013). *The impact of national and international assessment programmes on education policy, particularly policies regarding resource allocation and teaching and learning practices in developing countries*. Retrieved from http://research.acer.edu.au/ar_misc/16
Białecki, I., Jakubowski, M., & Wiśniewski, J. (2017). Education policy in Poland: The impact of PISA (and other international studies). *European Journal of Education, 52*(2), 167-174.
Black, P., & Wiliam, D. (2007). Large-scale assessment systems: Design principles drawn from international comparisons. *Measurement: Interdisciplinary Research and Perspectives, 5*(1), 1-53.
Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*(1), 5-31.
Bloem, S. (2015). PISA for low- and middle-income countries. *Compare: A Journal of Comparative and International Education, 45*(3), 481-486.
Bock, R. D. (1997). A Brief History of Item Response Theory. *Educational Measurement: Issues and Practice, 16*(4), 21-33.
Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.
Bolsinova, M., de Boeck, P., & Tijmstra, J. (2017). Modelling Conditional Dependence Between Response Time and Accuracy. *Psychometrika / Online preview*.
Bolsinova, M., & Tijmstra, J. (2016). Posterior Predictive Checks for Conditional Independence Between Response Time and Accuracy. *Journal of Educational and Behavioral Statistics, 41*(2), 123-145.
Bolsinova, M., Tijmstra, J., Molenaar, D., & De Boeck, P. (2017). Conditional Dependence between Response Time and Accuracy: An Overview of its Possible Sources and Directions for Distinguishing between Them. *Frontiers in Psychology, 8*(202).
Bonnet, G. (2002). Reflections in a Critical Eye: on the pitfalls of international assessment. *Assessment in Education: Principles, Policy and Practice, 9*(3), 387-399.
Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods, 1*(2), 97-111.
Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2014). Comprehensive Meta Analysis (CMA) (Version 3.3.070). Engelwood, NJ: Biostat.
Bradlow, E., T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64*(2), 153-168.
Braeken, J. (2011). A Boundary Mixture Approach to Violations of Conditional Independence. *Psychometrika, 76*(1), 57-76.
Brandt, S. (2008). Estimation of a Rasch model including subdimensions. In M. von Davier & D. Hastedt (Eds.), IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments (Vol. 1, pp. 51-69). Princeton: NJ: IEA-ETS Research Institute.
Brandt, S. (2010). Estimating tests including subtests. *Journal of Applied Measurement, 11*(4), 352-367.
Brandt, S. (2012). Robustness of multidimensional analyses against local item dependence. *Psychological Test and Assessment Modeling, 54*(1), 36-53.
Breakspear, S. (2012). *The Policy Impact of PISA: An Exploration of the Normative Effects of International Benchmarking in School System Performance. OECD Education Working Papers, No. 71*. Paris: OECD Publishing.
Brennan, R. L. (2006a). *Educational measurement* (4th ed.). Westport, CT: Praeger Publishers.
Brennan, R. L. (2006b). Perspectives on the Evolution and Future of Educational Measurement. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.). Westport, CT: Praeger.

Brown, G., Micklewright, J., Schnepf, S. V., & Waldmann, R. (2007). International surveys of educational achievement: How robust are the findings? *Journal of the Royal Statistical Society. Series A: Statistics in Society, 170*(3), 623-646.
Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (Second edition. ed.). New York: The Guilford Press.
Buysse, D. J., Yu, L., Moul, D. E., Germain, A., Stover, A., Dodds, N. E., . . . Pilkonis, P. A. (2010). Development and validation of patient-reported outcome measures for sleep disturbance and sleep-related impairments. *Sleep, 33*(6), 781-792.
Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika, 75*(4), 581-612.
Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized Full-Information Item Bifactor Analysis. *Psychological Methods, 16*(3), 221-248.
Cao, Y., Lu, R., & Tao, W. (2014). *Effect of Item Response Theory (IRT) Model Selection on Testlet-Based Test Equating. ETS Research Report (RR-14-19)*: Wiley Periodicals, Inc. Retrieved from https://www.ets.org/research/policy_research_reports/publications/report/2014/jsww
Cappelleri, J. C., Jason Lundy, J., & Hays, R. D. (2014). Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. *Clinical Therapeutics, 36*(5), 648-662.
Care, E., Griffin, P., Zhang, Z., & Hutchinson, D. (2014). Large-scale testing and its contribution to learning. In C. Wyatt-Smith, V. Klenowski, & P. Colbert (Eds.), *Designing Assessment for Quality Learning* (pp. 55-71). Dordrecht, Netherlands: Springer.
Carvalho, L. M., & Costa, E. (2015). Seeing education with one's own eyes and through PISA lenses: considerations of the reception of PISA in European countries. *Discourse: Studies in the Cultural Politics of Education, 36*(5), 638-646.
Carvalho, L. M., Costa, E., & Gonçalves, C. (2017). Fifteen years looking at the mirror: On the presence of PISA in education policy processes (Portugal, 2000-2016). *European Journal of Education, 52*(2), 154-166.
Chen, C. T., & Wang, W. C. (2007). Effects of ignoring item interaction on item parameter estimation and detection of interacting items. *Applied Psychological Measurement, 31*(5), 388-411.
Chen, J. (2014). *Model selection for IRT equating of Testlet-based tests in the random groups design.* (PhD), University of Iowa, Retrieved from http://ir.uiowa.edu/etd/1439
Chen, S. K., Hwang, F. M., & Lin, S. S. J. (2013). Satisfaction Ratings of QOLPAV: Psychometric Properties Based on the Graded Response Model. *Social Indicators Research, 110*(1), 367-383.
Chen, T. A. (2012). Fixed or random testlet effects: a comparison of two multilevel testlet models. *Journal of Applied Measurement, 13*(3), 231-247.
Chen, W.-H., & Thissen, D. (1997). Local Dependence Indexes for Item Pairs Using Item Response Theory. *Journal of Educational and Behavioral Statistics, 22*(3), 265-289.
Chou, Y.-T., & Wang, W.-C. (2010). Checking Dimensionality in Item Response Models With Principal Component Analysis on Standardized Residuals. *Educational and Psychological Measurement, 70*(5), 717-731.
Christensen, K. B. (2012). Ask the Experts: Rasch vs. Factor Analysis. *Rasch Measurement Transactions, 26*(3), 1373-1378.
Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical Values for Yen's Q3: Identification of Local Dependence in the Rasch Model Using Residual Correlations. *Applied Psychological Measurement, 41*(3), 178-194.
Chung, H., Kim, J., Cook, K. F., Askew, R. L., Revicki, D. A., & Amtmann, D. (2014). Testing measurement invariance of the patient-reported outcomes measurement information system pain behaviors score between the US general population sample and a sample of individuals with chronic pain. *Quality of Life Research, 23*(1), 239-244.

Cole, J. C., Motivala, S. J., Khanna, D., Lee, J. Y., Paulus, H. E., & Irwin, M. R. (2005). Validation of single-factor structure and scoring protocol for the Health Assessment Questionnaire-Disability Index. *Arthritis Care and Research, 53*(4), 536-542.

Cook, K. F., Kallen, M. A., & Amtmann, D. (2009). Having a Fit: Impact of Number of Items and Distribution of Data on Traditional Criteria for Assessing IRT's Unidimensionality Assumption. *Quality of Life Research, 18*(4), 447-460.

Cook, K. F., Teal, C. R., Bjorner, J. B., Cella, D., Chang, C.-H., Crane, P. K., . . . Reeve, B. B. (2007). IRT health outcomes data analysis project: an overview and summary. *Quality of Life Research, 16*(1), 121-132.

Cosgrove, J., & Cartwright, F. (2014). Changes in achievement on PISA: the case of Ireland and implications for international assessment practice. *Large-scale Assessments in Education, 2*(1), 2.

Crane, P. K., Carle, A., Gibbons, L. E., Insel, P., Mackin, R. S., Gross, A., . . . Mungas, D. (2012). Development and assessment of a composite score for memory in the Alzheimer's Disease Neuroimaging Initiative (ADNI). *Brain Imaging and Behavior, 6*(4), 502-516.

Crins, M. H. P., Roorda, L. D., Smits, N., de Vet, H. C. W., Westhovens, R., Cella, D., . . . Terwee, C. B. (2016). Calibration of the Dutch-Flemish PROMIS Pain Behavior item bank in patients with chronic pain. *European Journal of Pain, 20*(2), 284-296.

Crocker, L. M., & Algina, J. (2008). *Introduction to classical and modern test theory*. Ohio ; United Kingdom: Wadsworth.

d'Agnese, V. (2015). PISA's colonialism: Success, money, and the eclipse of education. *Power and Education, 7*(1), 56-72.

Davis, S. F., Drinan, P. F., & Bertram Gallant, T. (2009). *Cheating in school: what we know and what we can do*. Malden, MA: Wiley-Blackwell.

Davis, S. F., Drinan, P. F., & Bertram Gallant, T. (2017). Q & A with the authors of "Cheating in school : what we know and what we can do". Retrieved from http://au.wiley.com/WileyCDA/PressRelease/pressReleaseId-56517,descCd-release_additional_material.html

De Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: Guilford Press.

De Champlain, A. F. (2010). A primer on classical test theory and item response theory for assessments in medical education. *Medical Education, 44*(1), 109-117.

de Klerk, M., Nel, J. A., Hill, C., & Koekemoer, E. (2013). The development of the MACE work-family enrichment instrument. *SA Journal of Industrial Psychology, 39*(2), 1-16.

Debeer, D., & Janssen, R. (2013). Modeling item-position effects within an IRT framework. *Journal of Educational Measurement, 50*(2), 164-185.

Debelak, R., & Arendasy, M. (2012). An Algorithm for Testing Unidimensionality and Clustering Items in Rasch Measurement. *Educational and Psychological Measurement, 72*(3), 375-387.

DeMars, C. E. (2006). Application of the Bi-Factor Multidimensional Item Response Theory Model to Testlet-Based Tests. *Journal of Educational Measurement, 43*(2), 145-168.

DeMars, C. E. (2013). A comparison of confirmatory factor analysis and multidimensional Rasch models to investigate the dimensionality of test-taking motivation. *Journal of Applied Measurement, 14*(2), 179-196.

DEPP. (2007). *L'évaluation internationale PISA 2003: compétences des élèves francais en mathématiques, compréhension de l'écrit et Sciences*. Les Dossiers. Paris, France: MEN.

Dimitriadou, C., Gakoudi, A., Kalaitzidou-Leontaki, A., & Kousaridis, K. (2012). Students' Trends and Attitudes on Exam Cheating in Greek Primary and Secondary School Settings. In D. Alt & R. Reingold (Eds.), *Changes in Teachers' Moral Role: From Passive Observers to Moral and Democratic Leaders* (pp. 31-44). Rotterdam: SensePublishers.

DiStefano, C., & Morgan, G. B. (2014). A Comparison of Diagonal Weighted Least Squares Robust Estimation Techniques for Ordinal Data. *Structural Equation Modeling, 21*(3), 425-438.

Domínguez, M., Vieira, M. J., & Vidal, J. (2012). The impact of the Programme for International Student Assessment on academic journals. *Assessment in Education: Principles, Policy and Practice, 19*(4), 393-409.
Douglas, J., Kim, H. R., Habing, B., & Gao, F. (1998). Investigating Local Dependence With Conditional Covariance Functions. *Journal of Educational and Behavioral Statistics, 23*(2), 129-151.
Eckes, T. (2014). Examining testlet effects in the TestDaF listening section: A testlet response theory modeling approach. *Language Testing, 31*(1), 39-61.
Elley, W. B. (2005). How Timss-R Contributed to Education in Eighteen Developing Countries. *Prospects, 35*(2), 199-212.
Engelhard, G., Jr. (2012). Ask the Experts: Rasch vs. Factor Analysis. *Rasch Measurement Transactions, 26*(3), 1373-1378.
Ensoy, C., Rakhmawati, T. W., Faes, C., & Aerts, M. (2015). *Separation Issues and Possible Solutions: Part I – Systematic Literature Review on Logistic Models - Part II – Comparison of different methods for separation in logistic regression* (EN-869). Retrieved from http://onlinelibrary.wiley.com/doi/10.2903/sp.efsa.2015.EN-869/abstract
Epskamp, S., Cramer, A. O., Waldorp, L. J., Schmittmann, V. D., & Borsboom, D. (2012). Qgraph: Network visualizations of relationships in psychometric data. *Journal of Statistical Software, 48*(4), 1-18.
Ercikan, K., Chen, M. Y., Lyons-Thomas, J., Goodrich, S., Sandilands, D., Roth, W.-M., & Simon, M. (2015). Reading Proficiency and Comparability of Mathematics and Science Scores for Students From English and Non-English Backgrounds: An International Perspective. *International Journal of Testing, 15*(2), 153-175.
eResearch South Australia. (2017). High-performance computing. Retrieved from https://www.ersa.edu.au/service/hpc/
Erosheva, E. A., Fienberg, S. E., & Junker, B. W. (2002). Alternative statistical models and representations for large sparse multi-dimensional contingency tables. *Annales de la Faculté des sciences de Toulouse: Mathématiques, 11*(4), 485-505. Retrieved from http://www.numdam.org/item?id=AFST_2002_6_11_4_485_0
Fayers, P. M. (2007). Applying item response theory and computer adaptive testing: the challenges for health outcomes assessment. *Quality of Life Research, 16*(1), 187-194.
Feniger, Y., & Lefstein, A. (2014). How not to reason with PISA data: an ironic investigation. *Journal of Education Policy, 29*(6), 845-855.
Ferrara, S., Huynh, H., & Baghi, H. (1997). Contextual Characteristics of Locally Dependent Open-Ended Item Clusters in a Large-Scale Performance. *Applied Measurement in Education, 10*(2), 123-144.
Ferrara, S., Huynh, H., & Michaels, H. (1999). Contextual Explanations of Local Dependence in Item Clusters in a Large Scale Hands-On Science Performance Assessment. *Journal of Educational Measurement, 36*(2), 119-140.
Ferrer, A. T. (2017). PISA in Spain: Expectations, impact and debate. *European Journal of Education, 52*(2), 184-191.
Finch, W. H. (2012). The MIMIC Model as a Tool for Differential Bundle Functioning Detection. *Applied Psychological Measurement, 36*(1), 40-59.
Finch, W. H., & Jeffers, H. (2016). A Q3-Based Permutation Test for Assessing Local Independence. *Applied Psychological Measurement, 40*(2), 157-160.
Finney, S. J., & DiStefano, C. (2013). Non-normal and categorical data in structural equation modelling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling : a second course* (2nd ed.). Charlotte, NC: Information Age Publishing, Inc.
Fladmoe, A. (2012). Education in the news and in the mind. PISA, news media and public opinion in Norway, Sweden and Finland. *Nordicom Review, 33*(1), 99-116.

Flens, G., Smits, N., Terwee, C. B., Dekker, J., Huijbrechts, I., & de Beurs, E. (2017). Development of a Computer Adaptive Test for Depression Based on the Dutch-Flemish Version of the PROMIS Item Bank. *Evaluation & the Health Professions, 40*(1), 79-105.

Frey, A., Hartig, J., & Rupp, A. A. (2009). An NCME instructional module on booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice, 28*(3), 39-53.

Frey, A., & Seitz, N. N. (2009). Multidimensional adaptive testing in educational and psychological measurement: Current state and future challenges. *Studies in Educational Evaluation, 35*(2-3), 89-94.

Frey, A., & Seitz, N. N. (2011). Hypothetical use of multidimensional adaptive testing for the assessment of student achievement in the programme for international student assessment. *Educational and Psychological Measurement, 71*(3), 503-522.

Fukuhara, H., & Paek, I. (2016). Exploring the Utility of Logistic Mixed Modeling Approaches to Simultaneously Investigate Item and Testlet DIF on Testlet-based Data. *Journal of Applied Measurement, 17*(1), 79-90.

Garrido, L. E., Abad, F. J., & Ponsoda, V. (2016). Are fit indices really fit to estimate the number of factors with categorical variables? Some cautionary findings via monte carlo simulation. *Psychological Methods, 21*(1), 93-111.

Gebhardt, E., & Adams, R. J. (2007). The influence of equating methodology on reported trends in PISA. *Journal of Applied Measurement, 8*(3), 305-322.

Gilmore, A. (2005). The impact of PIRLS (2001) and TIMSS (2003) in low-and middle-income countries: An evaluation of the value of World Bank support for international surveys of reading literacy (PIRLS) and mathematics and science (TIMSS). Retrieved from http://pub.iea.nl/fileadmin/user_upload/Publications/Electronic_versions/Gilmore_Impact_PIRLS_TIMSS.pdf

Goldstein, H. (2004). International comparisons of student attainment: Some issues arising from the PISA study. *Assessment in Education, 11*(3), 319.

Gonzalez, E., & Rutkowski, L. (2010). Principles of multiple matrix booklet designs and parameter recovery in large-scale assessments. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments, 3*, 125-156.

Gorur, R. (2011). ANT on the PISA Trail: Following the statistical pursuit of certainty. *Educational Philosophy and Theory, 43*(sup1), 76-93.

Gorur, R., & Wu, M. L. (2015). Leaning too far? PISA, policy and Australia's 'top five' ambitions. *Discourse: Studies in the Cultural Politics of Education, 36*(5), 647-664.

Grace-Martin, K. (2017). Logistic Regression Models for Multinomial and Ordinal Variables. Retrieved from http://www.theanalysisfactor.com/logistic-regression-models-for-multinomial-and-ordinal-variables/

Grisay, A., & Griffin, P. (2006). Where are the main cross-national studies? In K. N. Ross & I. J. Genevois (Eds.), *Cross-national studies of the quality of education: planning their design and managing their impact* (pp. 67-103). Paris: UNESCO.

Grubbs, F. E. (1950). Sample Criteria for Testing Outlying Observations. *Annals of Mathematical Statistics, 21*(1), 27-58.

Guilford, J. P. (1936). *Psychometric Methods*. New York: McGraw-Hill.

Gustafsson, J.-E., & Rosen, M. (2006). The dimensional structure of reading assessment tasks in the IEA reading literacy study 1991 and the Progress in International Reading Literacy Study 2001. *Educational Research and Evaluation, 12*(5), 445 - 468.

Habing, B., & Roussos, L. A. (2003). On the need for negative local item dependence. *Psychometrika, 68*(3), 435-451.

Haley, S. M., Fragala-Pinkham, M. A., Dumas, H. M., Ni, P., Gorton, G. E., Watson, K., . . . Tucker, C. A. (2009). Evaluation of an item bank for a computerized adaptive test of activity in children with cerebral palsy. *Physical Therapy, 89*(6), 589-600.

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. Boston, MA: Kluwer Academic Publishers.
Hamilton, C. B., Maly, M. R., Giffin, J. R., Clark, J. M., Speechley, M., Petrella, R. J., & Chesworth, B. M. (2015). Validation of the Questionnaire to Identify Knee Symptoms (QuIKS) using Rasch analysis. *Health and Quality of Life Outcomes, 13*(157).
Hastedt, D., & Desa, D. (2015). Linking errors between two populations and tests: A case study in international surveys in education. *Practical Assessment, Research and Evaluation, 20*(14), 1-12.
Hays, R. D., Morales, L. S., & Reise, S. P. (2000). Item Response Theory and Health Outcomes Measurement in the 21st Century. *Medical Care, 38(9) Supplement*(II), II-28-II-42.
Hedberg, E. C. (2016). ICCVAR: Stata Module to Calculate Intraclass Correlation (ICC) after xtmixed. (Version Aug 2013): Boston College Department of Economics. Retrieved from https://ideas.repec.org/c/boc/bocode/s457468.html
Hedges, L. V., Hedberg, E. C., & Kuyper, A. M. (2012). The Variance of Intraclass Correlations in Three- and Four-Level Models. *Educational and Psychological Measurement, 72*(6), 893-909.
Heene, M., Hilbert, S., Freudenthaler, H. H., & Bühner, M. (2012). Sensitivity of SEM fit indexes with respect to violations of uncorrelated errors. *Structural Equation Modeling, 19*(1), 36-50.
Henning, G. (1989). Meanings and implications of the principle of local independence. *Language Testing, 6*(1), 95-108.
Hill, C. D., Edwards, M. C., Thissen, D., Langer, M. M., Wirth, R. J., Burwinkle, T. M., & Varni, J. W. (2007). Practical issues in the application of item response theory: A demonstration using items from the Pediatric Quality of Life Inventory (PedsQL) 4.0 generic core scales. *Medical Care, 45*(5 SUPPL. 1), S39-S47.
Hissbach, J. C., Klusmann, D., & Hampe, W. (2011). Dimensionality and predictive validity of the HAM-Nat, a test of natural sciences for medical school admission. *BMC Medical Education, 11*(83), 1-11.
Holt, t. J. C., van Duijn, M. A. J., & Boomsma, A. (2010). Scale construction and evaluation in practice: A review of factor analysis versus item response theory applications. *Psychological Test and Assessment Modeling, 52*(3), 272-297.
Hopfenbeck, T. N., & Görgen, K. (2017). The politics of PISA: The media, policy and public responses in Norway and England. *European Journal of Education, 52*(2), 192-205.
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*, 1-21.
Hoskens, M., & De Boeck, P. (1997). A Parametric Model for Local Dependence Among Test Items. *Psychological Methods, 2*(3), 261-277.
Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York: Wiley.
Houts, C. R., & Edwards, M. C. (2015). Comparing Surface and Underlying Local Dependence Levels via Polychoric Correlations. *Applied Psychological Measurement, 39*(4), 293-302.
Huang, H. Y., & Wang, W. C. (2013). Higher Order Testlet Response Models for Hierarchical Latent Traits and Testlet-Based Items. *Educational and Psychological Measurement, 73*(3), 491-511.
Huynh, H., & Ferrara, S. (1994). A Comparison of Equal Percentile and Partial Credit Equatings for Performance-Based Assessments Composed of Free-Response Items. *Journal of Educational Measurement, 31*(2), 125-141.
Huynh, H., Michaels, H., & Ferrara, S. (1995). *Statistical Procedures To Identify Clusters Of Items With Local Dependency.* Paper presented at the National Council on Measurement in Education, San Francisco.
IBM Corp. (2015). IBM SPSS Statistics for Windows (Version 23.0.0.3). Armonk, NY: IBM Corp.

Ikeda, T. (2015). Applying PISA Ideas to Classroom Teaching of Mathematical Modelling. In K. Stacey & R. Turner (Eds.), *Assessing mathematical literacy : The PISA experience.* (pp. 221-238): Cham: Springer.
Ip, E. H. (2001). Testing for local dependency in dichotomous and polytomous item response models. *Psychometrika, 66*(1), 109-132.
Ip, E. H. (2010). Interpretation of the Three-Parameter Testlet Response Model and Information Function *Applied Psychological Measurement, 34*(7), 467-482.
James, M. (2010). An overview of educational assessment. In *International Encyclopedia of Education. Vol. 3* (pp. 161-171). Oxford: Elsevier.
Jerrim, J. (2013). The reliability of trends over time in international education test scores: Is the performance of England's secondary school pupils really in relative decline? *Journal of Social Policy, 42*(2), 259-279.
Jiao, H., Kamata, A., Wang, S., & Jin, Y. (2012). A Multilevel Testlet Model for Dual Local Dependence. *Journal of Educational Measurement, 49*(1), 82-100.
Jiao, H., Wang, S., & Kamata, A. (2005). Modeling Local Item Dependence with the Hierarchical Generalized Linear Model. *Journal of Applied Measurement, 6*(3), 311-321.
Johnson, S. (1999). International Association for the Evaluation of Educational Achievement Science Assessment in Developing Countries. *Assessment in Education: Principles, Policy & Practice, 6*(1), 57-73.
Jones, R. N., Tommet, D., Ramirez, M., Jensen, R., & Teresi, J. A. (2016). Differential item functioning in Patient Reported Outcomes Measurement Information System® (PROMIS®) Physical Functioning short forms: Analyses across ethnically diverse groups. *Psychological Test and Assessment Modeling, 58*(2), 371-402
Kalpakjian, C. Z., Tulsky, D. S., Kisala, P. A., & Bombardier, C. H. (2015). Measuring grief and loss after spinal cord injury: Development, validation and psychometric characteristics of the SCI-QOL grief and loss item bank and short form. *Journal of Spinal Cord Medicine, 38*(3), 347-355.
Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling, 15*(1), 136-153.
Kamens, D H., & McNeely, C. L. (2010). Globalization and the Growth of International Educational Testing and National Assessment. *Comparative Education Review, 54*(1), 5-25.
Kanes, C., Morgan, C., & Tsatsaroni, A. (2014). The PISA mathematics regime: knowledge structures and practices of the self. *Educational Studies in Mathematics, 87*(2), 145-165.
Kaplan, D. (1995). Statistical Power in Structural Equation Modelling. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications*. Thousand Oaks, CA: Sage Publications.
Kappenburg -ten Holt, J. (2014). *A comparison between factor analysis and item response theory modeling in scale analysis.* (PhD), University of Groningen, Groningen: Netherlands.
Kaspar, R., & Hartig, J. (2016). Emotional competencies in geriatric nursing: empirical evidence from a computer based large scale assessment calibration study. *Advances in Health Sciences Education, 21*(1), 105-119.
Kasper, D., Ünlü, A., & Gschrey, B. (2014). Sensitivity Analyses for the Mixed Coefficients Multinomial Logit Model. In M. Spiliopoulou, L. Schmidt-Thieme, & R. Janning (Eds.), *Data Analysis, Machine Learning and Knowledge Discovery* (pp. 389-396). Cham: Springer International Publishing.
Kell, M., & Kell, P. (2014a). Global Testing: PISA, TIMSS and PIRLS. In *Literacy and Language in East Asia* (Vol. 24, pp. 33-49). Singapore: Springer.
Kell, M., & Kell, P. (2014b). International Testing: The Global Education Space Race? In *Literacy and Language in East Asia* (Vol. 24, pp. 51-77). Singapore: Springer.
Kellaghan, T., & Greaney, V. (2001). The globalisation of assessment in the 20th century. *Assessment in Education, 8*(1), 87-102.

Keller, L. A., Swaminathan, H., & Sireci, S. G. (2003). Evaluating scoring procedures for context-dependent item sets. *Applied Measurement in Education, 16*(3), 207-222.
Kelley, T. L. (1924). Note on the reliability of a test: A reply to Dr. Crum's criticism. *The Journal of Educational Psychology, 15*, 193-204.
Kent, P., Grotle, M., Dunn, K. M., Albert, H. B., & Lauridsen, H. H. (2015). Rasch analysis of the 23-item version of the Roland Morris Disability Questionnaire. *Journal of Rehabilitation Medicine, 47*(4), 356-364.
Kim, D. (2007). *Assessing the relative performance of local item dependence indexes.* (PhD), The University Of Nebraska, Lincoln.
Kim, D., De Ayala, R. J., Ferdous, A. A., & Nering, M. L. (2011). The Comparative Performance of Conditional Independence Indices. *Applied Psychological Measurement, 35*(6), 447–471.
Kim, J., Chung, H., Amtmann, D., Revicki, D. A., & Cook, K. F. (2013). Measurement invariance of the PROMIS pain interference item bank across community and clinical samples. *Quality of Life Research, 22*(3), 501-507.
Kisala, P. A., Tulsky, D. S., Kalpakjian, C. Z., Heinemann, A. W., Pohlig, R. T., Carle, A., & Choi, S. W. (2015). Measuring anxiety after spinal cord injury: Development and psychometric characteristics of the SCI-QOL Anxiety item bank and linkage with GAD-7. *Journal of Spinal Cord Medicine, 38*(3), 315-325.
Kline, R. B. (2013). Exploratory and confirmatory factor analysis. In Y. M. Petscher, C. Schatschneider, & D. L. Compton (Eds.), *Applied quantitative analysis in education and the social sciences* (pp. 171-207). New York: Routledge.
Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). New York: The Guilford Press.
Koğar, E. Y., & Kelecioğlu, H. (2017). Examination of Different Item Response Theory Models on Tests Composed of Testlets. *Journal of Education and Learning, 6*(4), 113-126.
Kreiner, S. (2011). *Is the foundation under PISA solid? A critical look at the scaling model underlying international comparisons of student attainment. Research Report 11/1*: Department of Biostatistics, University of Copenhagen. Retrieved from http://ls.ncm.gu.se/media/ncm/dokument/pisa_kreiner_.pdf
Kreiner, S., & Christensen, K. B. (2004). Analysis of Local Dependence and Multidimensionality in Graphical Loglinear Rasch Models. *Communications in Statistics - Theory and Methods, 33*(6), 1239 - 1276.
Kreiner, S., & Christensen, K. B. (2011). Item Screening in Graphical Loglinear Rasch Models. *Psychometrika, 76*(2), 228-256.
Kreiner, S., & Christensen, K. B. (2013a). *Assessment of item correlation in Rasch models*. Paper presented at the SIS 2013 Statistical Conference - Advances in Latent Variables - Methods, Models and Applications, Department of Economics and Management, Brescia. http://meetings.sis-statistica.org/index.php/sis2013/ALV/paper/viewFile/2558/458
Kreiner, S., & Christensen, K. B. (2013b). Two Tests of Local Independence. In K. B. Christensen, S. Kreiner, & M. Mesbah (Eds.), *Rasch Models in Health* (pp. 131-136). Hoboken, NJ: John Wiley & Sons.
Kreiner, S., & Christensen, K. B. (2014). Analyses of Model Fit and Robustness. A New Look at the PISA Scaling Model Underlying Ranking of Countries According to Reading Literacy. *Psychometrika, 79*(2), 210-231.
Kuha, J. (2004). AIC and BIC: Comparisons of assumptions and performance. *Sociological Methods and Research, 33*(2), 188-229.
Kuittinen, S., García Velázquez, R., Castaneda, A. E., Punamäki, R.-L., Rask, S., & Suvisaari, J. (2017). Construct validity of the HSCL-25 and SCL-90-Somatization scales among Russian, Somali and Kurdish origin migrants in Finland. *International Journal of Culture and Mental Health, 10*(1), 1-18.
Kwak, C., & Clayton-Matthews, A. (2002). Multinomial logistic regression. *Nursing research, 51*(6), 404-410.

Lawson, D. M., & Brailovsky, C. (2006). The Presence and Impact of Local Item Dependence on Objective Structured Clinical Examinations Scores and the Potential Use of the Polytomous, Many-Facet Rasch Model. *Journal of Manipulative and Physiological Therapeutics, 29*(8), 651-657.

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction* (pp. 362-412). Princeton, NJ: Princeton University Press.

Le Hebel, F., Montpied, P., & Tiberghien, A. (2014). Which Effective Competencies Do Students Use in PISA Assessment of Scientific Literacy? In C. Bruguière, A. Tiberghien, & P. Clément (Eds.), *Topics and Trends in Current Science Education* (Vol. 1, pp. 273-289): Springer Netherlands.

Le Hebel, F., Montpied, P., Tiberghien, A., & Fontanieu, V. (2017). Sources of difficulty in assessment: example of PISA science items. *International Journal of Science Education, 39*(4), 468-487.

Lee, G., Kolen, M. J., Frisbie, D. A., & Ankenmann, R. D. (2001). Comparison of Dichotomous and Polytomous Item Response Models in Equating Scores From Tests Composed of Testlets *Applied Psychological Measurement, 25*(4), 357-372.

Lee, Y.-W. (2004). Examining passage-related local item dependence (LID) and measurement construct using Q3 statistics in an EFL reading comprehension test. *Language Testing, 21*(1), 74-100.

Levy, R. (2011). Posterior Predictive Model Checking for Conjunctive Multidimensionality in Item Response Theory. *Journal of Educational and Behavioral Statistics, 36*(5), 672-694.

Levy, R., Mislevy, R. J., & Sinharay, S. (2009). Posterior Predictive Model Checking for Multidimensionality in Item Response Theory. *Applied Psychological Measurement, 33*(7), 519-537.

Levy, R., & Svetina, D. (2011). A generalized dimensionality discrepancy measure for dimensionality assessment in multidimensional item response theory. *British Journal of Mathematical and Statistical Psychology, 64*(2), 208-232.

Levy, R., Xu, Y., Yel, N., & Svetina, D. (2015). A Standardized Generalized Dimensionality Discrepancy Measure and a Standardized Model-Based Covariance for Dimensionality Assessment for Multidimensional Models. *Journal of Educational Measurement, 52*(2), 144-158.

Lietz, P., Cresswell, J., Rust, K., & Adams, R. J. (2017). *Implementation of large-scale education assessments*. Hoboken, NJ: John Wiley & Sons.

Lietz, P., & Tobin, M. (2016). The impact of large-scale assessments in education on education policy: evidence from around the world. *Research Papers in Education, 31*(5), 499-501.

Linacre, J. M. (2009). Local independence and residual covariance: A study of olympic figure skating ratings. *Journal of Applied Measurement, 10*(2), 157-169.

Linden, W. J. v. d. (2016). *Handbook of item response theory. Volume One: Models*. Boca Raton, FL: Chapman and Hall/CRC.

Linn, R. L. (2010). Educational Measurement: Overview. In P. P. B. McGaw (Ed.), *International Encyclopedia of Education (Third Edition)* (pp. 45-49). Oxford: Elsevier.

Linn, R. L. (Ed.) (1989). *Educational measurement* (3rd ed.). New York: Macmillan.

Liu, X. (2016). *Applied ordinal logistic regression using Stata : from single-level to multilevel modeling*. Los Angeles: Sage.

Liu, Y., & Maydeu-Olivares, A. (2013). Local Dependence Diagnostics in IRT Modeling of Binary Data. *Educational and Psychological Measurement, 73*(2), 254-274.

Liu, Y., & Thissen, D. (2012). Identifying Local Dependence With a Score Test Statistic Based on the Bifactor Logistic Model. *Applied Psychological Measurement, 36*(8), 670-688.

Liu, Y., & Thissen, D. (2014). Comparing score tests and other local dependence diagnostics for the graded response model. *British Journal of Mathematical and Statistical Psychology, 67*(3), 496-513.
Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores (with contributions by A. Birnbaum).* Reading, MA: Addison-Wesley.
Lundgren-Nilsson, A., Jonsdottir, I., Ahlborg, G., & Tennant, A. (2013). Construct validity of the psychological general well being index (PGWBI) in a sample of patients undergoing treatment for stress-related exhaustion: a rasch analysis. *Health and Quality of Life Outcomes, 11*(2).
Lyons-Thomas, J., Sandilands, D. D., & Ercikan, K. (2014). Gender differential item functioning in mathematics in four international jurisdictions. *Education and Science, 39*(172), 20-32.
Madaus, G. F., & O'Dwyer, L. M. (1999). A Short History of Performance Assessment: Lessons Learned. *Phi Delta Kappan, 80*(9), 688-695.
Mandíková, D., & Bašátková, K. (2008). *Analýza dat z mezinárodních výzkumů – fyzikální úlohy Úlohy výzkumu PISA* Univerzity Karlovy v Praze - Prague. Retrieved from https://kdf.mff.cuni.cz/vyzkum/NPVII/materialy/zprava_ulohy_pisa_komplet.pdf
Manrique-Vallier, D., & Fienberg, S. E. (2008). Population size estimation using individual level mixture models. *Biometrical Journal, 50*(6), 1051-1063.
Marais, I. (2013). Local Dependence. In *Rasch Models in Health* (pp. 111-130). Hoboken, NJ: John Wiley & Sons.
Marais, I., & Andrich, D. (2008a). Effects of Varying Magnitude and Patterns of Response Dependence in the Unidimensional Rasch Model. *Journal of Applied Measurement, 9*(1), 105-124.
Marais, I., & Andrich, D. (2008b). Formalizing dimension and response violations of local independence in the unidimensional Rasch model. *Journal of Applied Measurement, 9*(3), 200-215.
Martens, K., & Niemann, D. (2013). When do numbers count? The differential impact of the PISA rating and ranking on education policy in Germany and the US. *German Politics, 22*(3), 314-332.
Martin, M. O., Mullis, I. V. S., & Chrostowski, S. J. (Eds.). (2004). *TIMSS 2003 technical report: findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades*. Chestnut Hill, Mass.: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.
Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149-174.
Masters, G. N., & Forster, M. (2000). *The assessments we need*. Australian Council for Educational Research, Camberwell, Vic. Retrieved from http://cunningham.acer.edu.au/inted/theassessmentsweneed.pdf
Mattsson, M., Fearghal, O., Lajunen, T., Gormley, M., & Summala, H. (2015). Measurement invariance of the Driver Behavior Questionnaire across samples of young drivers from Finland and Ireland. *Accident Analysis and Prevention, 78*, 185-200.
Maydeu-Olivares, A. (2015). Evaluating the Fit of IRT Models. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling : applications to typical performance assessment* (pp. 111-128). New York, NY: Routledge.
Maydeu-Olivares, A., Cai, L., & Hernández, A. (2011). Comparing the fit of item response theory and factor analysis models. *Structural Equation Modeling, 18*(3), 333-356.
Maydeu-Olivares, A., & Joe, H. (2014). Assessing Approximate Fit in Categorical Data Analysis. *Multivariate Behavioral Research, 49*(4), 305-328.
Mazzeo, J., & von Davier, M. (2008). *Review of the Programme for International Student Assessment (PISA) test design: recommendations for fostering stability in assessment results*. Retrieved from http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=EDU/PISA/GB%282008%2928&docLanguage=En

McDonald, R. P. (1979). The Structural Analysis of Multivariate Data: A Sketch of a General Theory. *Multivariate Behavioral Research, 14*(1), 21-38.
McDonald, R. P. (1981). The Dimensionality of Tests and Items. *British Journal of Mathematical and Statistical Psychology, 34*, 100-117.
McDonald, R. P. (1982). Linear Versus Models in Item Response Theory. *Applied Psychological Measurement, 6*(4), 379-396.
McGaw, B. (2008). The role of the OECD in international comparative studies of achievement. *Assessment in Education: Principles, Policy & Practice, 15*(3), 223-243.
MedCalc Software BVBA. (2017). MedCalc Statistical Software (Version 17.6 (15 days licence)). Ostend, Belgium: MedCalc Software bvba.
Meyer, H.-D., & Benavot, A. (2013). *PISA, power, and policy : the emergence of global educational governance*. Didcot, UK: Symposium Books.
Michel, A. (2017). The contribution of PISA to the convergence of education policies in Europe. *European Journal of Education, 52*(2), 206-216.
Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology, 88*(3), 355-383.
Milford, T., Ross, S. P., & Anderson, J. O. (2010). An opportunity to better understand schooling: The growing presence of PISA in the Americas. *International Journal of Science and Mathematics Education, 8*(3), 453-473.
Mislevy, R. J. (1991). Randomization-Based Inference about Latent Variables from Complex Samples. *Psychometrika, 56*(2), 177-196.
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*(2), 133-161.
Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling Procedures in NAEP. *Journal of Educational Statistics, 17*(2), 131-154.
Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus Article: On the Structure of Educational Assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*(1), 3-62.
Moineddin, R., Matheson, F. I., & Glazier, R. H. (2007). A simulation study of sample size for multilevel logistic regression models. *BMC Medical Research Methodology, 7*(34), 1-10.
Monseur, C., & Adams, R. J. (2009). Plausible values: How to deal with their limitations. *Journal of Applied Measurement, 10*(3), 320-334.
Monseur, C., Baye, A., Lafontaine, D., & Quittre, V. (2011). PISA test format assessment and the local independence assumption. In M. von Davier & D. Hastedt (Eds.), *IERI monograph series - Issues and methodologies in large scale assessment* (Vol. IV, pp. 131-156). Hamburg, Germany: IEA/ETS Research Institute (IERI).
Monseur, C., & Berezner, A. (2007). The computation of equating errors in international surveys in education. *Journal of Applied Measurement, 8*(3), 323-335.
Muraki, E. (1992). A Generalized Partial Credit Model: Application of an EM Algorithm. *Applied Psychological Measurement, 16*(2), 159-176.
Muthén, L. K., & Muthén, B. O. (1998-2015). *Mplus User's Guide. Seventh Edition.* Los Angeles, CA: Muthén & Muthén.
Muthén, L. K., & Muthén, B. O. (2007-2017). Mplus Discussion - Empty Bivariate Table is not equal to 1? http://www.statmodel.com/discussion/messages/9/2496.html?1395581302
Muthén, L. K., & Muthén, B. O. (2015). Mplus (Version 7.4). Los Angeles, CE: Muthén & Muthén.
Nag, A. K., Mitra, A., & Mitra, S. (2005). Multiple outlier detection in multivariate data using self-organizing maps title. *Computational Statistics, 20*(2), 245-264.
Nandakumar, R., & Ackerman, T. (2004). Test modeling. In D. Kaplan (Ed.), *The SAGE handbook of quantitative methodology for the social sciences*. Thousand Oaks, Calif.: Sage.
Natesan, P., & Kieftenbeld, V. (2013). Measuring Urban Teachers' Beliefs About African American Students: A Psychometric Analysis. *Journal of Psychoeducational Assessment, 31*(1), 3-15.

National Center for Education Statistics. (2016a). Mathematics Literacy Items and Scoring Guides (2006 and 2012). Retrieved from http://nces.ed.gov/surveys/pisa/pdf/items2_math2012.pdf
National Center for Education Statistics. (2016b). PISA Released Paper-Based Assessment (PBA) Items and Scoring Guides. Retrieved from http://nces.ed.gov/surveys/pisa/educators.asp
National Center for Education Statistics. (2016c). Reading Literacy Items and Scoring Guides (2000 and 2009). Retrieved from http://nces.ed.gov/surveys/pisa/pdf/items2_reading.pdf
Naumann, J. (2005). TIMSS, PISA, PIRLS and Low Educational Achievement in World Society *Prospects, 5*(2), 229-248.
NCSS LLC. (2017). NCSS 10 Statistical Software (Version 10.0.14). Kaysville, Utah, USA: NCSS, LLC.
Neidorf, T. S., Binkley, M., Gattis, K., & Nohara, D. (2006). *Comparing Mathematics Content in the National Assessment of Educational Progress (NAEP), Trends in International Mathematics and Science Study (TIMSS), and Program for International Student Assessment (PISA) 2003 Assessments (NCES 2006-029).* Washington, DC: U.S. Department of Education, National Center for Education Statistics.
Niss, M. (2015). Mathematical Competencies and PISA. In K. Stacey & R. Turner (Eds.), *Assessing Mathematical Literacy: The PISA Experience* (pp. 35-55). Cham: Springer International Publishing.
Nye, C. D., & Drasgow, F. (2011). Assessing goodness of fit: Simple rules of thumb simply do not work. *Organizational Research Methods, 14*(3), 548-570.
OECD. (1999). *Measuring Student Knowledge and Skills: A New Framework for Assessment*. Paris: OECD Publishing.
OECD. (2001). *Knowledge and Skills for Life: First Results from PISA 2000*. Paris: OECD Publishing.
OECD. (2002). *Sample Tasks from the PISA 2000 Assessment Reading, Mathematical and Scientific Literacy*. Paris: OECD Publishing.
OECD. (2004). *The PISA 2003 Assessment Framework: Mathematics, Reading, Science and Problem Solving Knowledge and Skills*. Paris: OECD Publishing.
OECD. (2005a). *PISA 2003 Data Analysis Manual: SAS Users*. Paris: OECD Publishing.
OECD. (2005b). *PISA 2003 Technical Report*. Paris: OECD Publishing.
OECD. (2006). *Assessing Scientific, Reading and Mathematical Literacy: A Framework for PISA 2006*. Paris: OECD Publishing.
OECD. (2007a). Annex A3. In OECD (Ed.), *PISA 2006: Science Competencies for Tomorrow's World: Volume 1: Analysis*. Paris: OECD Publishing.
OECD. (2007b). *DEVELOPMENT OF THE PISA 2009 FIELD TRIAL INSTRUMENTS - 24th meeting of the PISA Governing Board - EDU/PISA/GB(2007)31*. Edinburgh, UK: OECD Publishing. Retrieved from http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=EDU/PISA/GB(2007)31&docLanguage=En
OECD. (2008). *ISSUES ARISING FROM THE PISA 2009 FIELD TRIAL PAPER-AND-PENCIL COGNITIVE ASSESSMENT - 26th meeting of the PISA Governing Board - EDU/PISA/GB(2007)31*. Hague: OECD Publishing.
OECD. (2009a). CODER RECRUITMENT KIT ADMINSTRATION GUIDELINES. Retrieved from https://www.acer.edu.au/files/coder_recruitment_guidelines_pisa09_2.pdf
OECD. (2009b). *PISA 2006 Technical Report*. Paris: OECD Publishing.
OECD. (2009c). *PISA Data Analysis Manual: SPSS, Second Edition*. Paris: OECD Publishing.
OECD. (2009d). *Take the Test Sample Questions from OECD's PISA Assessments: Sample Questions from OECD's PISA Assessments*. Paris: OECD Publishing.
OECD. (2010a). *PISA 2009 Assessment Framework: Key Competencies in Reading, Mathematics and Science*. Paris: OECD Publishing.
OECD. (2010b). *PISA Learning Mathematics for Life: A Perspective from PISA*. Paris: OECD Publishing.

OECD. (2012). *PISA 2009 Technical Report*. Paris: OECD Publishing.
OECD. (2013). *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy*. Paris: OECD Publishing.
OECD. (2014a). ANNEX A6. In OECD (Ed.), PISA 2012 Results: What Students Know and Can Do (Volume I, Revised edition, February 2014): Student Performance in Mathematics, Reading and Science. Paris: OECD Publishing. Retrieved from https://www.oecd.org/pisa/keyfindings/Annex_A6_Vol1.pdf.
OECD. (2014b). *PISA 2012 Technical Report*. Paris: OECD Publishing.
OECD. (2015a). *PISA Products. Codebooks for PISA 2000 (versions current as of 15th of Jan 2015)*. Retrieved from: https://www.oecd.org/pisa/pisaproducts/
OECD. (2015b). *PISA Products. Codebooks for PISA 2003 (versions current as of 15th of Jan 2015)*. Retrieved from: https://www.oecd.org/pisa/pisaproducts/
OECD. (2015c). *PISA Products. Codebooks for PISA 2006 (versions current as of 15th of Jan 2015)*. Retrieved from: https://www.oecd.org/pisa/pisaproducts/
OECD. (2015d). *PISA Products. Codebooks for PISA 2009 (versions current as of 15th of Jan 2015)*. Retrieved from: https://www.oecd.org/pisa/pisaproducts/
OECD. (2015e). *PISA Products. Databases for PISA 2000, 2003, 2006, 2009, 2012 (versions current as of 15th of Jan 2015)*. Retrieved from: https://www.oecd.org/pisa/pisaproducts/
OECD. (2016a). *PISA 2015 Results (Volume I): Excellence and Equity in Education*. Paris: OECD Publishing.
OECD. (2016b). PISA Test Questions. Retrieved from https://www.oecd.org/pisa/pisaproducts/pisa-test-questions.htm
OECD, & UNESCO Institute for Statistics. (2003). *Literacy skills for the world of tomorrow: further results from PISA 2000*. Paris: OECD Publishing.
Oliden, P. E., & Lizaso, J. M. (2013). Invariance levels across language versions of the PISA 2009 reading comprehension tests in Spain. *Psicothema, 25*(3), 390-395.
Oliden, P. E., & Lizaso, J. M. (2014). Impact of family language and testing language on reading performance in a bilingual educational context. *Psicothema, 26*(3), 328-335.
Oliveri, M. E., & von Davier, M. (2014). Toward Increasing Fairness in Score Scale Calibrations Employed in International Large-Scale Assessments. *International Journal of Testing, 14*(1), 1-21.
Paek, I., Yon, H., Wilson, M., & Kang, T. (2009). Random parameter structure and the testlet model: Extension of the Rasch testlet model. *Journal of Applied Measurement, 10*(4), 394-407.
Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology, 49*(12), 1373-1379.
Penfield, R. D. (2014). An NCME instructional module on polytomous item response theory models. *Educational Measurement: Issues and Practice, 33*(1), 36-48.
Pommerich, M., & Segall, D. O. (2008). Local dependence in an operational CAT: Diagnosis and implications. *Journal of Educational Measurement, 45*(3), 201-223.
Pons, X. (2017). Fifteen years of research on PISA effects on education governance: A critical review. *European Journal of Education, 52*(2), 131-144.
Popkewitz, T. (2011). PISA. In M. A. Pereyra, H.-G. Kotthoff, & R. Cowen (Eds.), *PISA Under Examination*. Dordrecht: Springer.
Prais, S. J. (2003). Cautions on OECD's recent educational survey (PISA). *Oxford Review of Education, 29*(2), 139-163.
Prais, S. J. (2004). Cautions on OECD's recent educational survey (PISA): Rejoinder to OECD's response. *Oxford Review of Education, 30*(4), 569-573.
Prenzel, M., Kobarg, M., Schöps, K., & Rönnebeck, S. (2013). *Research on PISA: research outcomes of the PISA research conference 2009*. New York: Springer.

Punch, K. (2014). *Introduction to social research : quantitative & qualitative approaches* (3rd ed.). Los Angeles, California: SAGE.
Randall, J., & Engelhard Jr, G. (2010). Using confirmatory factor analysis and the Rasch model to assess measurement invariance in a high stakes reading assessment. *Applied Measurement in Education, 23*(3), 286-306.
Ravand, H. (2015). Assessing Testlet Effect, Impact, Differential Testlet, and Item Functioning Using Cross-Classified Multilevel Measurement Modeling. *SAGE Open, 5*(2).
Reese, L. M. (1995). *The impact of local dependencies on some LSAT outcomes (LSAC research report 95-02)*. Newtown, PA: Law School Admission Council.
Reese, L. M. (1999). *Impact of Local Item Dependence on Item Response Theory Scoring in CAT. (LSAC research report 98-08)*. Princeton, NJ: Law School Admission Council.
Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., . . . Cella, D. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care, 45*(5 SUPPL. 1), S22-S31.
Reise, S. P., & Revicki, D. A. (2015). *Handbook of item response theory modeling : applications to typical performance assessment.* New York: Routledge, Taylor & Francis Group.
Resnik, L., Tian, F., Ni, P., & Jette, A. (2012). Computer-adaptive test to measure community reintegration of Veterans. *Journal of Rehabilitation Research and Development, 49*(4), 557-566.
Robitaille, D. F., & Beaton, A. E. (2002). TIMSS: A Brief Overview of The Study. In D. Robitaille & A. Beaton (Eds.), *Secondary Analysis of the TIMSS Data* (pp. 11-18). Dordrecht, Netherlands: Kluwer Academic Publishers.
Rodriguez, H. P., & Crane, P. K. (2011). Examining multiple sources of differential item functioning on the clinician & group CAHPS survey. *Health Services Research, 46*(6 PART 1), 1778-1802.
Rosner, B. (2011). *Fundamentals of biostatistics* (7th ed.). Boston: Brooks/Cole, Cengage Learning.
Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.
Ruddock, G., Clausen-May, T., Purple, C., & Ager, R. (2006). *Validation study of the PISA 2000, PISA 2003 and TIMSS-2003 international studies of pupil attainment* (Research Report 772): National Foundation for Educational Research. Retrieved from http://webarchive.nationalarchives.gov.uk/20130321032807/https://www.education.gov.uk/publications/eOrderingDownload/RR772.pdf
Rutkowski, L., Gonzales, E., von Davier, M., & Zhou, Y. (2013). Assessment Design for International Large-Scale Assessments. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis* (pp. 75-95). Boca Raton, FL: Taylor & Francis.
Rutkowski, L., & Rutkowski, D. (2016). A Call for a More Measured Approach to Reporting and Interpreting PISA Results. *Educational Researcher, 45*(4), 252-257.
Rutkowski, L., Rutkowski, D., & Zhou, Y. (2016). Item Calibration Samples and the Stability of Achievement Estimates and System Rankings: Another Look at the PISA Model. *International Journal of Testing, 16*(1), 1-20.
Salonen, A. H., Rosenström, T., Edgren, R., Volberg, R., Alho, H., & Castrén, S. (2017). Dimensions of the South Oaks Gambling Screen in Finland: A cross-sectional population study. *Scandinavian Journal of Psychology, 58*(3), 228-237.
Salzberger, T. (2012). Ask the Experts: Rasch vs. Factor Analysis. *Rasch Measurement Transactions, 26*(3), 1373-1378.
Savalei, V. (2011). What to do about zero frequency cells when estimating polychoric correlations. *Structural Equation Modeling, 18*(2), 253-273.
Schleicher, A. (2017). Seeing education through the prism of PISA. *European Journal of Education, 52*(2), 124-130.

Schreiber, J. B., Stage, F. K., King, J., Nora, A., & Barlow, E. A. (2006). Reporting Structural Equation Modeling and Confirmatory Factor Analysis results: A review. *Journal of Educational Research, 99*(6), 323-337.
Serder, M., & Ideland, M. (2016). PISA truth effects: the construction of low performance. *Discourse: Studies in the Cultural Politics of Education, 37*(3), 341-357.
Shiel, G., & Eivers, E. (2009). International comparisons of reading literacy: What can they tell us? *Cambridge Journal of Education, 39*(3), 345-360.
Sireci, S. (2015). Beyond ranking of nations: Innovative research on PISA. *Teachers College Record, 117*(1), 1-8.
Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the Reliability of Testlet-Based Tests. *Journal of Educational Measurement, 28*(3), 237-247.
Skaggs, G. (2007). Bookmark Locations and Item Response Model Selection in the Presence of Local Item Dependence. *Journal of Applied Measurement, 8*(1), 65-83.
Smith, E. V., Jr. (2005). Effects of item redundancy on Rasch item and person estimates. *Journal of Applied Measurement, 6*(2), 147-163.
Smith, R. M. (1996). A comparison of methods for determining dimensionality in Rasch measurement. *Structural Equation Modeling, 3*(1), 25-40.
Smits, N., Zitman, F., Cuijpers, P., den Hollander-Gijsman, M., & Carlier, I. (2012). A proof of principle for using adaptive testing in routine Outcome Monitoring: the efficiency of the Mood and Anxiety Symptoms Questionnaire -Anhedonic Depression CAT. *BMC Medical Research Methodology, 12*(1), 4.
Snowden, A., Watson, R., Stenhouse, R., & Hale, C. (2015). Emotional Intelligence and Nurse Recruitment: Rasch and confirmatory factor analysis of the trait emotional intelligence questionnaire short form. *Journal of Advanced Nursing, 71*(12), 2936-2949.
Sörbom, D. (1989). Model modification. *Psychometrika, 54*, 371-384.
Soussi, A., Broi, A.-M., Moreau, J., & Wirthner, M. (2004). *La littéracie dans quatre pays francophones: les résultats des jeunes de 15 ans en compréhension de l'écrit*. Neuchâtel, Switzerland: IRDP.
Soussi, A., Broi, A.-M., Moreau, J., & Wirthner, M. (2013). *La littératie en Suisse romande - PISA 2009: qu'en est-il des compétences des jeunes romands de 11eH, neuf ans après la première enquête?* Neuchâtel, Switzerland: IRDP.
Spray, J. A., & Ackerman, T. A. (1987). *The Effect of Item Response Dependency on Trait or Ability Dimensionality. (ACT Research Report 87-10)*. Iowa City, IA: American College Testing. Retrieved from http://209.235.214.158/research/researchers/reports/pdf/ACT_RR87-10.pdf
Stacey, K., & Turner, R. (2015a). *Assessing mathematical literacy: The PISA experience.*: Cham: Springer.
Stacey, K., & Turner, R. (2015b). The Evolution and Key Concepts of the PISA Mathematics Frameworks. In K. Stacey & R. Turner (Eds.), *Assessing Mathematical Literacy: The PISA Experience* (pp. 5-33). Cham: Springer.
StataCorp. (2015). *STATA MULTILEVEL MIXED-EFFECTS REFERENCE MANUAL Release 14*. College Station, TX: Stata Press.
StataCorp. (2017). Stata Statistical Software: Release 14 (Version 14.2). College Station, TX: StataCorp LP.
Steinberg, L., & Thissen, D. (1996). Uses of Item Response Theory and the Testlet Concept in the Measurement of Psychopathology. *Psychological Methods, 1*(1), 81-97.
Stone, C. A., & Yeh, C. C. (2006). Assessing the dimensionality and factor structure of multiple-choice exams: An empirical comparison of methods using the Multistate Bar Examination. *Educational and Psychological Measurement, 66*(2), 193-214.
Strietholt, R., & Rosén, M. (2016). Linking Large-Scale Reading Assessments: Measuring International Trends Over 40 Years. *Measurement: Interdisciplinary Research and Perspectives, 14*(1), 1-26.

Sulkunen, S. (2007). *Text authenticity in international reading literacy assessment. Focusing on PISA 2000.* (PhD), University of Jyväskylä, Jyväskylä, Finland. Retrieved from https://jyx.jyu.fi/dspace/handle/123456789/13434
Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*(3), 393-408.
Tasaki, N. (2017). The impact of OECD-PISA results on Japanese educational policy. *European Journal of Education, 52*(2), 145-153.
Tate, R. (2003). A Comparison of Selected Empirical Methods for Assessing the Structure of Responses to Test Items. *Applied Psychological Measurement, 27*(3), 159-203.
Teker, G. T., & Dogan, N. (2015). The effects of testlets on reliability and differential item functioning. *Educational Sciences: Theory & Practice, 15*(4), 969-980.
Teo, T. E. (2013). *Handbook of Quantitative Methods for Educational Research*. Rotterdam, Netherlands: Sense Publishers.
Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace Lines for Testlets: A Use of Multiple-Categorical-Response Models. *Journal of Educational Measurement, 26*(3), 247-260.
Thomas, M. L. (2011). The value of item response theory in clinical assessment: A review. *Assessment, 18*(3), 291-307.
Thompson, T. D., & Pommerich, M. (1996, April 8-12). *Examining the Sources and Effects of Local Dependence.* Paper presented at the annual meeting of the American Educational Research Association, New York.
Thorndike, R. L. (Ed.) (1951). *Educational Measurement*. Washington, DC: American Council on Education.
Thorndike, R. L., Angoff, W. H., Lindquist, E. F., & American Council on Education. (1971). *Educational Measurement* (2d ed.). Washington, DC: American Council on Education.
Transparency International. (2017). CORRUPTION PERCEPTIONS INDEX 2016. Retrieved from https://www.transparency.org/whatwedo/publication/corruption_perceptions_index_2016
Traub, R. E. (1997). Classical Test Theory in Historical Perspective. *Educational Measurement: Issues and Practice, 16*(4), 8-14.
Trendtel, M., Ünlü, A., Kasper, D., & Stubben, S. (2014). Using Latent Class Models with Random Effects for Investigating Local Dependence. In M. Spiliopoulou, L. Schmidt-Thieme, & R. Janning (Eds.), *Data Analysis, Machine Learning and Knowledge Discovery* (pp. 407-416). Cham: Springer International Publishing.
Tsai, T. H., Chaimongkol, S., & Hsu, Y. C. (2006). *Assessing local item dependence in the polytomously scored items.* Paper presented at the 37th Annual Conference of the Northeastern Educational Research Association (NERA), Kerhonkson, NY.
Tuerlinckx, F., & De Boeck, P. (2001a). The Effect of Ignoring Item Interactions on the Estimated Discrimination Parameters in Item Response Theory. *Psychological Methods, 6*(2), 181-195.
Tuerlinckx, F., & De Boeck, P. (2001b). Non-modeled item interactions lead to distorted discrimination parameters: A case study. *Methods of Psychological Research Online, 6*(2), 159-174.
Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley Publishing Company.
Turner, R., & Adams, R. J. (2007). The programme for international student assessment: An overview. *Journal of Applied Measurement, 8*(3), 237-248.
van der Lans, R. M., van de Grift, W. J. C. M., & van Veen, K. (2017). Developing an Instrument for Teacher Feedback: Using the Rasch Model to Explore Teachers' Development of Effective Teaching Strategies and Behaviors. *Journal of Experimental Education*, 1-18.
van der Linden, W. J., & Glas, C. A. W. (2010). Statistical tests of conditional independence between responses and/or response times on test items. *Psychometrika, 75*(1), 120-139.
van der Linden, W. J., Veldkamp, B. P., & Carlson, J. E. (2004). Optimizing Balanced Incomplete Block Designs for Educational Assessments. *Applied Psychological Measurement, 28*(5), 317-331.

van der Wollenberg, A. L. (1982). Two new test statistics for the Rasch model. *Psychometrika, 47*, 123-140.
van Rijn, P., & Rijmen, F. (2015). On the explaining-away phenomenon in multivariate latent variable models. *British Journal of Mathematical and Statistical Psychology, 68*(1), 1-22.
Vartanian, T. P. (2011). *Secondary data analysis*. New York: Oxford University Press.
Volante, L. (2013). Canadian Policy Responses to International Comparison Testing. *Interchange, 44*(3), 169-178.
von Davier, M. (2013). *The role of international large-scale assessments: perspectives from technology, economy, and educational research*. Dordrecht, Germany: Springer.
von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful? *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments, 2*, 9–36.
Wagemaker, H. (2004). *IEA: International Studies, Impact and Transition (Key Note).* Paper presented at the 1st IEA International Research Conference, Lefkosia, Cyprus.
Wagner, D. A. (2010). Quality of education, comparability, and assessment choice in developing countries. *Compare: A Journal of Comparative and International Education, 40*(6), 741-760.
Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. New York: Cambridge University Press.
Wainer, H., & Kiely, G. L. (1987). Item Clusters and Computerized Adaptive Testing: A Case for Testlets. *Journal of Educational Measurement, 24*(3), 185-201.
Wainer, H., & Thissen, D. (1996). How Is Reliability Related to the Quality of Test Scores? What Is the Effect of Local Dependence on Reliability? *Educational Measurement: Issues and Practice, 15*(1), 22-29.
Wainer, H., & Wang, X. (2000). Using a New Statistical Model for Testlets to Score TOEFL. *Journal of Educational Measurement, 37*(3), 203-220.
Waldow, F., Takayama, K., & Sung, Y. K. (2014). Rethinking the pattern of external policy referencing: Media discourses over the "Asian Tigers" PISA success in Australia, Germany and South Korea. *Comparative Education, 50*(3), 302-321.
Walter, O. B., & Rose, M. (2013). Effect of item order on item calibration and item bank construction for computer adaptive tests. *Psychological Test and Assessment Modeling, 55*(1), 81-91.
Wang, W.-C., Cheng, Y.-Y., & Wilson, M. (2005). Local Item Dependence for Items Across Tests Connected by Common Stimuli. *Educational and Psychological Measurement, 65*(1), 5-27.
Wang, Y. C., Hart, D. L., Deutscher, D., Yen, S. C., & Mioduski, J. E. (2013). Psychometric properties and practicability of the self-report Urinary Incontinence Questionnaire in patients with pelvic-floor dysfunction seeking outpatient rehabilitation. *Physical Therapy, 93*(8), 1116-1129.
Watt, T., Groenvold, M., Deng, N., Gandek, B., Feldt-Rasmussen, U., Rasmussen, A., . . . Bjorner, J. (2014). Confirmatory factor analysis of the thyroid-related quality of life questionnaire ThyPRO. *Health and Quality of Life Outcomes, 12*(1), 126.
Waugh, R. F., & Chapman, E. (2005). An Analysis of Dimensionality using Factor Analysis (True-Score Theory) and Rasch Measurement: What is the Difference? Which Method is Better? *Journal of Applied Measurement, 6*(1), 80-99.
Westfall, P. H., Henning, K. S. S., & Howell, R. D. (2012). The effect of error correlation on interfactor correlation in psychometric measurement. *Structural Equation Modeling, 19*(1), 99-117.
Wetzel, E., & Carstensen, C. H. (2013). Linking PISA 2000 and PISA 2009: Implications of instrument design on measurement invariance. *Psychological Test and Assessment Modeling, 55*(2), 181–206.

Williams, R. T., Heinemann, A. W., Bode, R. K., Wilson, C. S., Fann, J. R., & Tate, D. G. (2009). Improving Measurement Properties of the Patient Health Questionnaire-9 With Rating Scale Analysis. *Rehabilitation Psychology, 54*(2), 198-203.
Wilson, M., & Adams, R. J. (1995). Rasch Models for Item Bundles. *Psychometrika, 60*(2), 181-198.
Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods, 12*(1), 58-79.
Wiseman, A. W. (2010). *The impact of international achievement studies on national education policymaking*. Bingley, UK: Emerald.
Wu, M. L. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation, 31*(2-3), 114-128.
Wu, M. L. (2010). Measurement, Sampling, and Equating Errors in Large-Scale Assessments. *Educational Measurement: Issues and Practice, 29*(4), 15-27.
Wu, M. L., & Adams, R. J. (2002, 6-7 April). *Plausible Values – Why They Are Important.* Paper presented at the International Objective Measurement Workshop, New Orleans.
Wu, M. L., & Adams, R. J. (2013). Properties of rasch residual fit statistics. *Journal of Applied Measurement, 14*(4), 339-355.
Yang, K.-L., & Lin, F.-L. (2015). The effects of PISA in Taiwan: Contemporary Assessment Reform. In K. Stacey & R. Turner (Eds.), *Assessing mathematical literacy: The PISA experience.* (pp. 261-273): Cham: Springer.
Yen, W. M. (1984). Effects of Local Item Dependence on the Fit and Equating Performance of the Three-Parameter Logistic Model. *Applied Psychological Measurement, 8*(2), 125-145.
Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*(3), 187-213.
Yen, W. M., & Fitzpatrick, A. R. (2006). Item Response Theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111-153). Westport, CT: Praeger Publishers.
Yore, L. D., Anderson, J. O., & Chiu, M. H. (2010). Moving PISA results into the policy arena: Perspectives on knowledge transfer for future considerations and preparations. *International Journal of Science and Mathematics Education, 8*(3), 593-609.
Yu, C. Y. (2002). *Evaluating Cutoff Criteria of Model Fit Indices for Latent Variable Models with Binary and Continuous Outcomes.* (PhD), University of California, Los Angeles, CA. Retrieved from http://statmodel2.com/download/Yudissertation.pdf
Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2002). Identification and evaluation of local item dependencies in the medical college admissions test. *Journal of Educational Measurement, 39*(4), 291-309.
Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2003). *Effects of Local Item Dependence on the Validity of IRT Item, Test, and Ability Statistics. MCAT Monograph*: Association of American Medical Colleges. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.134.5458&rep=rep1&type=pdf
Zhang, J. (2007). *Dichotomous or polytomous model? Equating of testlet-based tests in light of conditional item pair correlations.* (PhD), University of Iowa,
Zickar, M. J., & Broadfoot, A. A. (2008). The partial revival of a dead horse? Comparing classical test theory and item response theory. In *Statistical and Methodological Myths and Urban Legends: Doctrine, Verity and Fable in the Organizational and Social Sciences* (pp. 37-59).
Zimmermann, S., Klusmann, D., & Hampe, W. (2016). Are Exam Questions Known in Advance? Using Local Dependence to Detect Cheating. *PLoS ONE, 11*(12), e0167545.