



**Comparability of Science Assessment across Languages:
The Case of PISA Science 2006**

Thesis submitted for the degree

Doctor of Philosophy in Education

Yasmine Hachem El Masri

St Anne's College

Supervisors: Jo-Anne Baird, Jane McNicholl

Advisor: David Andrich

January 2015

To the one who instilled in me the eagerness to reach for the stars

My Wonderful Father

To the one who never doubted I can reach the stars

My Wonderful Mother

To my adorable parents who supported me in this journey in every way possible

Hachem and Dina

ACKNOWLEDGEMENTS

This journey would have not been as enriching and thrilling if it was not for the many people I met throughout. I was so fortunate to be extremely well supported academically, logistically and emotionally.

I would like to express my heartfelt gratitude to my wonderful supervisors, Professor Jo-Anne Baird and Dr Jane McNicholl. Their patience, encouragement and immense knowledge were key motivations throughout my DPhil.

I owe a great debt of gratitude to my advisor, Professor David Andrich (University of Western Australia), for all the thought-provoking conversations we had. David has given me invaluable feedback on the analyses and provided constructive comments on several chapters of this thesis.

I would like to thank my examiners, Dr Therese Hopfenbeck (Oxford University) and Dr Anton Béguin (CITO; AQA) for agreeing to examine my thesis and for engaging with my work in a very stimulating and constructive manner.

Special thanks to Dr Joshua McGrane (University of Western Australia) for all the patience and assistance he has shown at the data entry stage. Many thanks to Dr Jon Twing and Dr Steve Strand for the constructive feedback on Chapters 4 and 6 of this thesis.

Many thanks to Dr Marit Kjærnsli (University of Oslo, Norway) and Imad Ababneh (National Center for Human Resources Development, Jordan) for providing access to the French and Arabic versions of the released PISA items respectively.

Many thanks to Dr Peter van Rijn (ETS Global, Amsterdam) for his assistance in understanding technical procedures of PISA.

Many thanks to OUCEA, St Anne's College and the Department of Education for providing me with financial support to attend workshops and conferences.

Special thanks to the people at the Oxford University Centre for Educational Assessment (OUCEA) for their constant emotional support. I thank OUCEA's academics (Daniel Caro, Jenny Lenkeit), our precious administrator (Joanne Hazell), visiting fellows (namely Professor Art Graesser and Nia Dowell, Dr Jim Tognolini), and last but not least, my colleagues Carol Brown, María Teresa Florèz, Natalie Usher, Agni Paramita, Diana Ng and Jessica Chan.

Completing this work would have been all the more difficult were it not for the support of my dearest friends - Julian Nowag, Nisrine Jaafar, David Menassa, Ioanna Aslanidou, Clementina Piazza and Rani Subassandran. You have made my student years in Oxford unforgettable. Also thanks to my rowing clubs (St Anne's and Academics) as well as the Linacre Ladies that Lift club (namely the amazing Shehzad Naqvi).

My deepest appreciation goes to my previous supervisor, Professor Barend Vlaardingerbroek (American University of Beirut) for his constant support of my academic growth.

A special mention for Mimi Kester, for the lifetime skills and support she has provided me with in my first two years of DPhil.

Special thanks to the Association for Educational Assessment-Europe (AEA-E) members and namely to Alistair Pollitt for the several enlightening conversations.

ABSTRACT

In this research, I investigated the extent to which language versions (English, French and Arabic) of the same science test were comparable in terms of item difficulty and demands. I used PISA science 2006 data from three countries (respectively, UK, France and Jordan). I argued that language was an intrinsic part of the scientific literacy construct, be it intended or not by the examiner. The tight relationship between the language element and the scientific knowledge makes the language variable inextricable from the construct. This argument has considerable implications on methodologies used to address this question. I also argued that none of the available statistical or qualitative techniques were capable of teasing out the language variable and answering the research question.

In this thesis, I adopted a critical evaluation and empirical methods, using literature from various fields (cognitive linguistics, psychology, measurement and science education) to analyse the test development and design procedures. In addition, I illustrated my claims with evidence from the technical reports and examples of released items. I adopted the same class of models employed in PISA, the Rasch model, as well as differential item functioning (DIF) techniques to address my question empirically.

General tests of fit suggested an overall good fit of the data to the model with eleven items out of 103 showing strong evidence of misfit. Various violations to the requirements of the Rasch model were highlighted. The DIF analysis indicated that 22% of the items showed bias in the selected countries, but bias balanced out at test level. Limitations of the DIF analysis to identify the source of bias were discussed. Qualitative approaches to investigating question demands were examined and issues with their usefulness in international settings were discussed. A way forward incorporating cognitive load theory and computational linguistics is proposed.

TABLE OF CONTENTS

List of Tables	XV
List of Figures.....	XVI
CHAPTER 1 INTRODUCTION	1
1.1 Background and rationale	3
1.2 Conceptual controversies.....	6
1.3 Purpose and overarching research question	9
1.4 Methodology.....	9
1.5 Significance	10
CHAPTER 2 Language and Science Assessments across Cultures	12
2.1 Is language part of the construct in PISA science assessments?.....	15
2.2 What is a construct and how are educational constructs developed?	15
2.2.1 PISA's scientific literacy framework and constructs	18
2.2.2 Reading and writing demands in PISA assessments	20
2.2.2.a Reading demands of PISA items	21
2.2.2.b Writing demands in open-ended responses	25
2.3 A closer look at PISA's scientific literacy framework and proficiency levels.....	27
2.3.1 The fundamental sense of science literacy.....	27
2.3.1.a Reading and writing in science	29
2.3.1.b The complexity of the scientific language	30
2.3.2 The language element in PISA's scientific literacy framework	33
2.3.3 Language expectations in proficiency levels of PISA science.....	35
2.4 Are science constructs comparable across languages?	38
2.4.1 Perception of time and linguistic relativism	39
2.4.2 Different languages: Different cognitive and linguistic demands	41
CHAPTER 3 Equivalence of Language Versions	51
3.1 Equivalent forms of an assessment under CTT	54
3.1.1 Parallel forms	55
3.1.2 Levels of equivalence	58
3.2 Equivalent forms of an assessment under LTT	60
3.2.1 Types or levels of linkage?	60
3.2.2 Linking language versions of the same assessment	62
3.3 Test adaptation	66
3.3.1 Options for test translation.....	67
3.3.1.a Application option	67
3.3.1.b Adaptation option.....	68
3.3.1.c Assembly option.....	70
3.4 Types of equivalence and of bias in adapted tests	71
3.4.1 Types of equivalence.....	71
3.4.2 Types of bias	74
3.5 The challenges of the test adaptation.....	76
3.5.1 Cultural and language differences affecting scores.....	76
3.5.1.a Construct equivalence	76
3.5.1.b Quality of communication in test administration.....	77
3.5.1.c Familiarity with test format.....	78
3.5.1.d Speededness	79
3.5.2 Technical issues, designs and methods	80
3.5.2.a The nature of the test	80

3.5.2.b	Selection and training of translator.....	81
3.5.2.c	The process of translation	83
3.5.2.d	Judgmental designs for adapting tests.....	84
3.5.2.e	Techniques for assessing equivalence at item and at test level.....	86
3.6	Factors affecting interpretation of results	89
3.6.1	Differences in curricula	89
3.6.2	Student motivation.....	91
3.6.3	Socioeconomic and political factors	91
3.7	Standards and guidelines for test adaptation.....	93
3.7.1	Standards for Educational and Psychological Testing	93
3.7.2	The ITC Guidelines for Test Adaptation	95
CHAPTER 4	The Rasch Model	99
4.1	Overview of the Rasch model – the formal stance.....	102
4.2	Invariance in Rasch modelling – the empirical stance.....	104
4.2.1	The Rasch model and fundamental measurement	105
4.2.2	The Rasch model: Detecting anomaly in the data	107
4.3	Measurement in social sciences.....	110
4.3.1	The nature of the variable – the ontological stance	112
4.3.2	Is the variable measureable?.....	113
4.3.3	Is measurement in social sciences possible and what does it mean?	116
4.4	Requirements of Rasch modelling.....	118
4.4.1	The sufficient statistic	120
4.4.2	The number of parameters	121
4.4.3	Unidimensionality of the construct	122
CHAPTER 5	PISA Tests and Sample Description.....	128
5.1	The PISA science test.....	129
5.1.1	PISA test format.....	130
5.1.2	PISA test design.....	131
5.2	PISA 2006 Scientific Literacy Framework	133
5.3	Development of cognitive items.....	137
5.4	Translation and adaption of PISA 2006	140
5.4.1	Development of English and French source versions.....	141
5.4.2	Double translation from two source languages	142
5.5	Sampling design and item calibration.....	143
5.6	Data Analysis	144
5.6.1	The mixed coefficient logit model	145
5.6.2	Population model	146
5.6.3	PISA item calibration.....	148
5.6.3.a	Conditioning variables.....	148
5.6.3.b	Plausible values in PISA.....	149
5.6.3.c	Dimensions in science	149
5.6.4	OECD results	150
5.7	Countries selected	151
CHAPTER 6	Rasch Analysis and Fit of Data to the Model.....	157
6.1	The Rasch model	159
6.2	Data 160	
6.2.1	Data retrieval and recoding.....	161
6.2.2	Initial Sample.....	162
6.2.3	Reduced sample	163
6.3	Precision of the model and the power to detect misfit	164

6.3.1	Precision of the model.....	165
6.3.2	Power to detect misfit and reliability of a test – the Person Separation Index (PSI)	165
6.3.3	Relating precision and power	167
6.3.4	Adjusted chi square value.....	168
6.3.5	Bonferroni correction	168
6.4	Targeting – Alignment of persons and item thresholds.....	170
6.5	Analysis of fit.....	172
6.5.1	Overall fit statistics.....	173
6.5.1.a	Item-person interaction.....	173
6.5.1.b	Item-trait interaction.....	174
6.5.2	Individual item-fit statistics.....	175
6.5.2.a	Fit residuals	175
6.5.2.b	Chi squares.....	177
6.5.2.c	Item characteristic curve (ICC).....	178
6.5.3	Results of individual item fit statistics	179
6.6	Random guessing.....	183
6.6.1	Impact of random guessing on item parameters	184
6.6.2	Impact of random guessing on person parameters	190
6.6.3	Conclusion.....	192
6.7	Local independence	193
6.7.1	Impact of violations to location independence on the unit of scale and reliability indices	195
6.7.2	Formalising local dependence	197
6.7.3	Analysis of response dependence	198
6.7.3.a	Dependent items	198
6.7.3.b	Magnitude of dependence.....	201
6.7.3.c	Conclusion	204
6.7.4	Analysis of trait dependence: Multidimensionality.....	204
6.7.4.a	Principal component analysis (PCA).....	205
6.7.4.b	Results.....	206
6.8	Order of thresholds	208
6.8.1	Order of thresholds in polytomous items of PISA Science 2006	209
CHAPTER 7	DIF and Distractor Analyses	217
7.1	Differential Item Functioning.....	220
7.2	DIF approach.....	228
7.2.1	DIF detection using ANOVA of response residuals.....	229
7.2.2	Detecting DIF: Artificial versus Real DIF.....	230
7.3	Results.....	233
7.3.1	DIF at item level.....	233
7.3.2	Quantifying DIF at item level.....	236
7.3.3	Was a particular country dis/advantaged overall?	238
7.3.3.a	Comparing means of DIF item location	239
7.3.3.b	Variation between person means before and after splitting	242
7.3.4	Conclusion of the DIF analyses.....	245
7.4	Distractor Analysis.....	246
7.4.1	Distractor proportion curves (DPCs)	247
7.4.2	Data	248
7.4.3	Results	250
7.4.3.a	Released simple MCQs flagged for DIF.....	250
7.4.3.b	Released simple MCQs not flagged for DIF	255

CHAPTER 8 Judging Test Demands across Languages	273
8.1 Difficulty versus demands	276
8.2 Types of demands	277
8.3 Approaches for judging question demands	280
8.3.1 Informal investigations of question demands	281
8.3.2 Systematic investigations of question demands	282
8.3.3 Construct elicitation technique	283
8.3.4 Systematic comparisons of language versions of PISA	285
8.4 Models for judging demands	287
8.4.1 Pollitt and Ahmed's Model of the Question Answering Process (QAP)	287
8.4.1.a Making Meaning While Reading	289
8.4.1.b Sources of easiness (SOEs) and sources of difficulty (SODs)	291
8.4.2 The CRAS framework	294
8.4.2.a Development of CRAS	294
8.4.2.b The CRAS scales of demands	295
8.4.3 Task difficulty	297
8.4.4 Challenges to judging demands	299
8.5 Problems with survey questions	304
8.6 Judging demands using cognitive load theory	308
8.6.1 Cognitive Load Theory	309
8.6.2 Effect of incongruence on the working memory (WM)	311
8.6.3 Techniques used to measure cognitive load	313
8.7 A closer look at some of the released items	314
8.7.1 Semiotics incongruence	314
8.7.2 Abbreviation incongruence	315
8.7.3 Interference between everyday language and scientific language	317
Chapter 9 Discussion	325
9.1 Translating meaning of popular idioms	327
9.2 Controversy 1: Language is embedded in what we are testing	329
9.3 Controversy 2: The statistical methods and models used in international assessments do not capture the complexity of educational constructs	333
9.3.1 Measuring or modelling?	334
9.3.2 Data fit to the model	335
9.3.3 The assumptions underlying IRT models	335
9.3.3.a Local independence	336
9.3.3.b Unidimensionality	336
9.4 Controversy 3: Available methods are not suitable for detecting invalidity of international comparisons	337
9.5 The way forward: Exploring cognitive science and computational linguistics	341
9.6 Limitations	343
9.7 Ethical issues	344
9.8 Concluding remarks	344
Reference List	347
APPENDICES	361
Appendix 3.1	362
Appendix 3.2 – ITC Guidelines for Test Adaptation	406
Appendix to 6.1 – Recoding PISA data in RUMM2030	408
Appendix to 7.1 – DIF items (Country curves)	411

GLOSSARY

AERA	American Educational Research Association
APA	American Psychological Association
CLT	cognitive load theory
CRAS	Complexity, Resources, Abstractness, task Strategy and response Strategy
CTT	classical test theory
DIF	differential item functioning
DPC	distractor proportion curve
ICC	item characteristic curve
IEA	International Association for the Evaluation of Educational Achievement
IRT	item response theory
ITC	International Testing Commission
LTT	latent test theory
MCQ	multiple choice question
NCME	National Council on Measurement in Education
OECD	Organisation for Economic Cooperation and Development
2-, 3-PL	2-, 3-parameter logistic (model)
PCA	principal component analysis
PIRLS	Progress in International Reading Literacy Study
PISA	Programme for International Student Assessment
QUAID	Question Understanding Aid
RMM	Reading Maturity Metric
RUMM	Rasch unidimensional measurement models
SOD	sources of difficulty
SOE	sources of easiness
TIMSS	Trends in International Maths and Science Study

List of Tables

Table 2.1: PISA 2006 scientific competencies	34
Table 3.1: Typical sources for three types of bias in cross-cultural assessment	74
Table 3.2: Designs for examining equivalence at test level.....	89
Table 5.1: Cluster rotation design used to form test booklets for PISA 2006	132
Table 5.2: Number and percentage of science cognitive items in terms of item format and competency.....	135
Table 5.3: Number and percentage of science cognitive items in terms of item format and knowledge type	136
Table 5.4: Number and percentage of science cognitive items in terms of knowledge category and competency	136
Table 5.5: Mean scores, standard errors and rankings of UK, France and Jordan on the science scale in PISA 2006.	153
Table 6.1: Proportions of countries in the initial sample and the reduced sample	164
Table 6.2: Item fit based on fit residuals and Chi Square tests after reducing sample to N=12,000 with equal country proportions.....	180
Table 6.3: Item locations after the tailored analysis and the anchored analyses of multiple choice items in increasing order of t-statistic	188
Table 6.4: Correlated items.....	200
Table 6.5: Individual item fit statistics for split items.....	202
Table 6.6: Magnitude of dependence d between correlated items:.....	203
Table 6.7 Summary of the principal component analysis	206
Table 6.8: Reversed order of thresholds.....	210
Table 6.9: Fit residuals and Chi Square test results before and after rescoring	213
Table 7.1: Studies comparing language versions of the same international test.....	224
Table 7.2: Two-way ANOVA of residuals from four class intervals	235
Table 7.3: Number of relatively dis/advantaging DIF items per country	239
Table 7.4: DIF item locations and average difficulty per country	240
Table 7.5: Within-subjects effects (repeated-measures ANOVA) at 5% significance level....	242
Table 7.6: Pairwise comparisons.....	242
Table 7.7: Person means before and after splitting with fixed mean	243
Table 7.8: Items eligible for distractor analysis.....	250
Table 8.1: The CRAS scales of demands.....	296
Table 8.2: Revised PISA reading item difficulty scheme.....	297
Table 8.3: Problems with survey questions	307

List of Figures

Figure 2.1: Scientific literacy framework of PISA 2006 science assessment	19
Figure 2.2: Released PISA science unit S195	23
Figure 2.3: TIMSS item S022154 (Grade 8)	24
Figure 2.4: PISA proficiency levels in scientific literacy	37
Figure 2.5: Which way is tomorrow in different cultures	40
Figure 2.6a: Moon Phases in different languages	43
Figure 2.6b: Same questions in three languages	43
Figure 3.1: Lego representation of the same construct in three languages	63
Figure 3.2: Lego representation of three constructs of the same underlying concept	64
Figure 3.3a: TIMSS 1995 Maths item (Population 2)	68
Figure 3.3b: Application of TIMSS 1995 Maths item (Population 2)	68
Figure 3.4a: TIMSS 1995 Maths item (Population 2)	69
Figure 3.4b: Adaptation of TIMSS 1995 Maths item (Population 2)	69
Figure 3.5a: PISA science 2006 multiple choice item	82
Figure 3.5b: Adaptation of item S508Q03 key into French and Arabic	83
Figure 3.6: Possible relationships between PISA and national curricula	91
Figure 5.1: Framework of PISA science 2006 assessment (OECD, 2006, p.26)	134
Figure 6.1a: Person-Item Threshold Location	171
Figure 6.1b: Person-Item Threshold Location per country	172
Figure 6.2: ICC of item i of difficulty δ_i	178
Figure 6.3: ICCs of items 10, 93 and 96 (N=12,000)	182
Figure 6.4: Anchored versus Tailored mean item locations	190
Figure 6.5: ICC of items 56, 73, 75, 91 and 96	192
Figure 6.6: Threshold map of polytomous items 32, 33, 34, 35, 37, 38	209
Figure 6.7a: Category probability curves of items 34 and 35	210
Figure 6.7b: Category probability curves of items 32, 33, 37 and 38	211
Figure 6.8: Threshold probability curves of polytomous items	212
Figure 7.1: Examples of DIF items (UK: blue, France: red, Jordan: green)	236
Figure 7.2: Country ICC for items 41, 43 and 59	237
Figure 7.3: Scatter plot matrix of pairs of countries based on locations of DIF items.	241
Figure 7.4: DPCs of Item 40 from the UK sample	248
Figure 7.5a: Simple MCQ (PISA Science 2006)	248
Figure 7.5b: Complex MCQ (PISA Science 2006)	249

Figure 7.6: Item locations per country for item 58	251
Figure 7.7: Distractor analysis of item 58	252
Figure 7.8: Item locations per country for item 59	252
Figure 7.9: Distractor analysis of item 59	253
Figure 7.10: Item locations per country for item 94	254
Figure 7.11: Distractor analysis of item 94	255
Figure 7.12: Item locations per country for item 40	256
Figure 7.13: Distractor analysis of item 40	257
Figure 7.14: Item locations per country for item 68	257
Figure 7.15: Distractor analysis of item 68	258
Figure 7.16: ICC curve of item 68 in Jordan	259
Figure 7.17: Item locations per country for item 69	260
Figure 7.18: Distractor analysis of item 69	261
Figure 7.19: Item locations per country for item 70	261
Figure 7.20: Distractor analysis of item 70	262
Figure 7.21: ICC curve of item 70 in France (left) and Jordan (right)	263
Figure 7.22: Item locations per country for item 80	263
Figure 7.23: Distractor analysis of item 80	264
Figure 7.24: Item locations per country for item 81	265
Figure 7.25: Distractor analysis of item 81	266
Figure 7.26: Item locations per country for item 85	266
Figure 7.27: Distractor analysis of item 85	267
Figure 8.1: TIMSS items	277
Figure 8.2: Relative amount of work per question	278
Figure 8.3: English and Arabic versions of graphs in PISA unit S114	315
Figure 8.4: Advertisements of cosmetic product in Arabic with the acronyms UVA and SPF in Latin letter	320
Figure 9.1: Quality of machine translation in 2015	341

FOREWORD

My interest in this research area stemmed from my experiences as a biology teacher in Lebanon where science instruction and assessment occur in English or French, even though Arabic is the first language. English was the medium of instruction in the school where I taught and our school had a sister school that provided the same lessons and administered the same exams in French. Being trilingual, translation of educational material and tests from English to French and vice versa has been an everyday practice. We, as teachers, rarely wondered whether the translation of the material had any impact on students' learning or achievement.

My experiences as a teacher are reflected in what happens at a larger scale in Lebanon where high-stakes examinations – The *Brevet Libanais* (age 15) and *Baccalauréat Libanais* (age 18) – are translated from French to English with very little quality control on the translation process. Comparability of results did not create much controversy amongst researchers, practitioners and users. There was an assumption that because the tests were 'genuinely' translated, they were of the same difficulty and demands and students had equal chances of success, irrespective of the language of the test. This of course might be true but there was no evidence to support that.

My initial DPhil project was to investigate whether the English and French versions of the same biology test of the *Baccalauréat Libanais* were equivalent in terms of item difficulty and demand. It built on my Master's thesis entitled *The Readability of the National Grade 12 Lebanese Biology Textbook: A Comparison between Anglophone and Francophone Learning Milieux* which compared the readability of language versions of the same science textbook in

Lebanon. However, my data collection was hampered by prolonged teacher strikes in July 2012 and a very unstable political situation in Lebanon since 2011. I therefore decided to raise the same research question at an international level. Fortunately, the new project has opened the door to many opportunities. I was introduced to Rasch modelling through two intensive courses with the University of Western Australia. I was hosted by Art Graesser as a visiting fellow at the Institute of Intelligent Systems (University of Memphis) to learn about applications of computational linguistics. I also won an internship at ETS Global Amsterdam and worked alongside Dr Peter van Rijn on a project using the same sample of PISA data I used in my DPhil analyses. These experiences gave me the opportunity to apply two- and three-parameter IRT models.

International large-scale assessments exposed me to deep and fundamental conceptual debates about measurement and the nature of the variable examined in educational measurement. They also allowed me to develop a critical stance about these issues as well as methodologies employed in this area.

CHAPTER 1

INTRODUCTION

‘يا مستعجل وقف تقلك.’

‘Hey you in a hurry! Stop so that I tell you.’

(Literal translation from Lebanese Arabic)

1.1 Background and rationale

Instances of comparing the achievement of students from different language backgrounds using the same test have recently become ever more on demand, with the rising importance of large-scale international assessments such as the Programme for International Student Assessment (PISA), the Trends in International Maths and Science Study (TIMSS) and the Progress in International Reading Literacy Study (PIRLS). PISA, governed by the Organisation for Economic Cooperation and Development (OECD), assesses literacy of 15-year old students in reading, mathematics and science. TIMSS and PIRLS are administered by the International Association for the Evaluation of Educational Achievement (IEA). TIMSS evaluates trends in mathematics and science achievement at fourth (Population 1) and eighth grade (Population 2) and PIRLS assesses trends in reading comprehension at fourth grade.

While international comparisons of student performances may be useful in understanding systemic and socio-cultural factors influencing education, the development and design of such tests has instigated much conceptual and technical controversy. The debates revolve around the possibility of developing a test for which language versions allow fair comparisons of student achievement across countries. An equitable comparison of scores across countries sitting the same test should be established on a premise that the various language forms are

equivalent for all participants and offer all parties equal opportunities of success. This means that the exams should be of comparable difficulty and demands, lest the assessment advantage or disadvantage a specific language group.

Despite the rigorous quality control exerted on the translation and adaptation processes in PISA and TIMSS, some items have been reported to behave differently in different countries (Ercikan & Koh, 2005; Grisay & Monseur, 2007; Grisay et al., 2007; Hauger & Sireci, 2008; Huang, Wilson & Wang, 2014; Kreiner & Christensen, 2014; Le, 2009; Oliveri & Ercikan, 2011; Sandilands et al., 2013; Wu & Ercikan, 2006; Xie & Wilson, 2008). With language being a culturally-laden, complex variable which promotes and influences thinking, devising equivalent science tests in different languages is a complicated endeavour. Indeed, in this thesis I argue that comparing language versions of the same science test is methodologically impossible. Language is an intrinsic part of the construct; i.e. what the test measures. Translation effects are unavoidable and hence bias, at some level, is inevitable. This threatens the validity and reliability of the tests. Available statistical techniques do not allow the identification of the sources of bias and qualitative techniques are not reliable. This makes it even more challenging to ensure comparability of scores across languages.

International tests are very costly for countries. Besides, their results have become very influential at policy level driving educational reforms in various countries (Ertl, 2006; Schleicher, 2009). It is hence important to ensure that these tests stand on solid conceptual and methodological grounds to justify both the allocation of national funds and educational reforms. A major problem in these tests is the impact of language on the extent to which they are comparable. In the context of international assessments, the lack of comparability has consequences at the policy and systemic levels. The tests bare low stakes for students, thus not

directly affecting their opportunities in life. However, in other contexts, such as in Lebanon, South Africa, Canada, Switzerland and Hong Kong, where more than one medium of instruction operates, the comparability of language versions of the same test could have tremendous impact on groups.

In Lebanon for instance, science education and assessment occur in English or French; not in the country's national language, Arabic. Similarly, high-stakes examinations also occur in English or French. They are administered at age 15 (*Brevet Libanais*) and at age 18 (*Baccalauréat Libanais*); the first serving as a bottleneck to secondary non-vocational education and the second acting mainly as a university entrance exam. Success in the *Brevet* and *Baccalauréat* determine opportunities in life for students sitting them. Nevertheless, translation of the tests is carried out with minimal quality control¹. Scores across language versions are directly compared with very few questions raised, if any, on the impact of translation on the demands and difficulty of questions. It is entirely assumed that careful translation of the tests would ensure their fairness.

Be it in the international or the national setting, comparison of student achievement across language versions of the same test is needed. However, the field lacks the appropriate methods to do so. Hence, exploring the conceptual framework underpinning these tests as well as shortcomings of the methodology adopted in developing them and analysing the data can provide tremendous insight for a way forward. Several studies have examined language effects in national (Ercikan et al., 2004; Jones, 1998) and international tests (Ercikan & Koh,

¹Based on an interview carried out with Dr Enja Osman on 12 April 2011. Dr Enja is a faculty member in the education department at the American University of Beirut. Having a long experience in the Lebanese Baccalaureate examinations, she has been a member of the examination committee for the life science subject for several years.

2005; Grisay & Monseur, 2007; Grisay et al., 2007; Hauger & Sireci, 2008; Huang et al., 2014; Oliveri & Ercikan, 2011; Sandilands et al., 2013). Some have looked at science tests specifically (Hauger & Sireci, 2008; Le, 2009; Huang et al., 2014). However, these studies explored biases as empirical outcomes and did not examine the conceptual controversies underlying the comparability of tests. Besides, none of these studies investigated the comparability of Arabic versions of tests. The following section provides an overview of the conceptual and methodological controversies explored in this thesis.

1.2 Conceptual controversies

Language is an important attribute of culture. In this thesis, culture refers to the standard definition of the Oxford dictionary; it is the customs and social behaviour of a particular group of individuals. Different cultures perceive the world differently and this is largely due to the environment in which they are embedded, the social norms and, according to the linguistic relativism principle, it is due to the language people speak (Carroll, 1956). Linguistic relativism theory relies on an intimate relationship between language and culture. The moderate version of this principle contends that language promotes and shapes thinking because of its lexical and syntactic characteristics. These lexical and syntactic features vary across languages and place different cognitive and linguistic demands on users, be it in speaking, reading or writing.

Differential language effects have great implications on international assessments. Assessments rely heavily on clear communication between examiners and examinees. This communication occurs mostly through oral or written language. Often, this communication imposes additional unintended demands on examinees turning language into a construct-

irrelevant variable (Oakland & Lane, 2004). However, language can be a part of the construct, if this component is measured in the test. In this case, differential language effects have an even greater significance in cross-country comparisons as the language component becomes even more salient. When measuring achievement in a particular subject across countries, it is important to make sure that demands associated with language do not advantage or disadvantage a particular group. The intimate relationship between culture and language is manifested in a very complex manner in international assessments. A main challenge is to tease the two variables out methodologically and distinguish their impact without implicating them together. Indeed, Chapter 3 describes the intricacies entailed in the translation and adaption of a given test into different languages and cultures.

Tremendous efforts are invested in translation and adaptation of international assessments (Grisay et al., 2007) to ensure equivalence of the language versions. However, based on the review of ample research carried out in translation and adaptation of instruments, bias in language versions seems inevitable even when rigorous procedures are followed (Hambleton et al., 2005). Bias could arise from the way constructs are manifested in the different cultures. It could also emerge from the design, methods and techniques employed in the process of translation and adaptation of tests or the way results are interpreted. Unfortunately until today, the evaluation of the quality of test translation and adaptation relies heavily on subjective expert judgement (OECD, 2009b; Arffman, 2012). These methods were found to yield inconsistent results (Pollitt et al, 2007). This raises questions about the validity and reliability of the tests and of international comparisons.

Another contentious aspect about international assessments is related to the use of item response theory (IRT) models to calibrate items and analyse data. Debates about whether

these models measure student abilities or model them come to the surface. While these models are powerful and allow direct comparison of scores across countries, they fail to capture the complexity of the educational reality (Goldstein, 1979). The assumptions underlying these models, such as unidimensionality and local independence, are often incongruent with the educational and curricular goals (Baird & Black, 2013; Goldstein, 1979). Besides, fierce criticism was expressed against the use of the Rasch model in PISA because of the great misfit of the data to the model (Goldstein, 2004; Kreiner & Christensen, 2014) and the large proportion of biased items (Kreiner & Christensen, 2014). These findings threaten the validity and reliability of PISA results and their country rankings.

Various statistical approaches of DIF are available to detect bias in items. Bias is reflected as significant difference in the difficulty of items across countries. DIF techniques are fairly simple to apply; however, they suffer from several limitations. They cannot detect bias when all items manifest it. Besides, they fail to identify the sources of bias which in the context of international tests could be associated with an endless number of variables: quality of translation and adaptation, nature of the construct, culture, language demands, educational systems, curricula, etc. Of particular importance to this research is the differential language demands which inevitably vary across languages according to the linguistic relativism principle. Available approaches and models for investigating demands placed on examinees during an assessment lack solid theoretical grounds and are incapable of predicting difficulty. Besides, they are not suitable for investigating differential demands across languages for reasons I outline in the thesis.

1.3 Purpose and overarching research question

The purpose of this research was to highlight the significance of language when comparing different language versions of the same test. The research attempted to answer the overarching question: do languages differentially impact the difficulty and demands of exams.

1.4 Methodology

In this research project I adopted a pragmatic stance where reality is considered to be ‘experiential [...] with different elements or layers, some objective, some subjective, and some a mixture of the two’ (Feilzer, 2010, p.8). This is reflected in my thesis in various places and namely in the methodology I adopted. I embraced a moderate version of linguistic relativism as a lens to examine the impact of language on thinking and rejected its deterministic stance. I also adopted the Rasch model to analyse the data based on pragmatic reasons and distanced myself from strong positivist claims related to measurement in social sciences that are often attributed to Rasch.

As a pragmatist, I espoused the view that dichotomy between quantitative and qualitative data is non-existent at an epistemological and ontological level (Tashakkori & Teddlie, 2003). I used both types of evidence to illustrate the conceptual and methodological controversies. I employed a critical evaluation of the data and assessment material (items, framework, technical manuals, etc.) by applying existing theories in different fields (educational assessment, science education, psychometrics, cognitive science, philosophy and linguistics) integrating them to see how they related to educational assessment. I used PISA 2006 science data from three countries (UK, France and Jordan) to illustrate the conceptual and

methodological controversies underpinning the research question. The thesis therefore used both critical evaluation and empirical techniques. It was thus highly conceptual with Chapters 2, 4 and 8 highlighting the theoretical debates encountered in this research and Chapters 5, 6 and 7 using data to illustrate the arguments I make.

1.5 Significance

This research aims to better understand challenges facing the comparability of language versions of the same assessment in a hope to improve the design, development and data analysis of cross-cultural and international assessments. It aims to contribute to both theory and practice in educational assessment.

CHAPTER 2

Language and Science Assessments across Cultures

Τα μάτια σου δεκατέσσερα. [Ta mátia sou dekatéssera]
Your eyes fourteen

(Literal translation from Greek)

Introduction

International comparisons of education systems often occur these days by means of administering translated versions of the same test. The assumption is that, with rigorous translation and adaptation techniques, constructs – that is, what is measured – could be translated into different languages. In other words, thorough translation and adaptation methods neutralise the potential language impact on constructs. However, empirical studies on English and French versions of the same TIMSS assessment indicated that constructs in mathematics and sciences are not necessarily equivalent across languages (Ercikan & Koh, 2005). The question worth asking here is whether language is part of the construct measured; or is it merely a construct-irrelevant variable.

In this chapter, I argue that language is part of the construct, irrespective of the intention of the question-setter. I use PISA science tests as a context to illustrate my argument. PISA builds its assessments using subject-specific frameworks (OECD, 1999). The framework adopted for the development of science assessments is one which focuses on ‘scientific literacy’; a common framework shaping science education around the world (OECD, 2006). The scientific literacy framework in PISA science defines three constructs which do not explicitly include language. They focus on acquiring scientific knowledge and science processing skills (e.g. critical thinking, formulating hypotheses, interpreting data, etc.).

However, looking closely at PISA assessments, a striking feature is the language density of the assessments and hence the reading demands it places on participants (Hopfenbeck & Maul, 2011). Also, open-ended questions require elaborate answers; thus introducing writing challenges. These aspects are likely to be manifested as having language as a construct-irrelevant variable, to say the least. Moreover, a review of the literature on scientific literacy suggests that reading, writing and speaking in science; that is, communication skills, are intertwined with more knowledge- and cognitive-based skills and thus they are essential to scientific literacy. This, hence, makes language part of the science construct in science assessments including PISA irrespective of the intention of the examiner. This conclusion is confirmed when inspecting the description of the performance levels of PISA where high performance in science is associated with good command of the scientific language skills.

This conclusion – that language is part of the construct in science assessments – has great implications on international comparisons. Hauger and Sireci (2008, p. 248) considered that:

“[i]t is difficult to establish that we are measuring the same construct with the same accuracy when we administer a test in a single language to different cultural groups. The problem becomes much trickier in international assessments that involve more than one language. When students from different countries respond to test questions in different languages, it is hard to establish construct equivalence [...]’.

Having language as part of the construct means that, even if we assume that the scientific element of the construct is not affected by the language used to express it, constructs differ across countries because of the language difference. Language is not a mere channel of thoughts; language promotes thinking and shapes it (Carroll, 1956). This relationship between language and thinking is emphasised in science. Scientists use specific linguistic features to express the precision and objectivity of their thinking (Fang, 2006; Moss, 2000;

Sutton, 1998). Although the scientific language used by scientists is different from the scientific language used in schools, both languages share similar characteristics (Halliday & Martin, 1993; Moss, 2000). These characteristics are very likely to permeate into science assessments. They are also likely to impose different cognitive demands in different languages. I argue here that this differential impact threatens the validity of international comparisons making them akin to ‘comparing apples and oranges’.

This chapter is divided into four sections. The first section examines the meaning of a construct and how constructs are produced in educational assessments. The second section presents the scientific framework adopted by PISA and examines the formulation of the competencies and the extent to which language is part of the constructs measured in PISA science. The second section investigates the importance of the language variable in PISA science assessments by first examining the language density in PISA science units and then inspecting the description of proficiency levels as defined in the PISA manuals. Although not explicit, I conclude that language is likely to constitute a construct-irrelevant variable in PISA science assessments.

Later, I take a closer look at the scientific framework and the competencies using the literature on scientific literacy and the science language. I argue that language is part of the construct in science assessments focusing on scientific literacy such as PISA, be it intended by the examiner or not. The last section investigates the extent to which science constructs are comparable across languages. Linguistic relativism is discussed in terms of the extent to which different languages determine different ways of thinking and hence confer different cognitive and linguistic challenges. Two examples are used to illustrate this point. The chapter ends with the discussion of the implications of such an argument on the validity of international comparisons using the same assessment in different languages.

2.1 Is language part of the construct in PISA science assessments?

The scientific literacy framework in PISA describes the scope, the content, the format and the constructs measured in PISA science tests. If language is part of the science construct, then it should appear as an aspect of the framework or it should be stated in the constructs. The following section describes PISA's scientific literacy framework and the constructs measured in the PISA science test.

2.2 What is a construct and how are educational constructs developed?

Before discussing whether language is part of the construct in science assessments, it is worth clarifying what a construct means and how constructs are produced in educational assessments. This section provides a definition of a construct and associates its importance to the validity of tests. It also briefly describes how constructs are defined in educational assessments.

Cronbach and Meehl (1955, p.283) defined a psychological construct as follows,

A construct is some postulated attribute of people, assumed to be reflected in test performance. In test validation the attribute about which we make statements in interpreting a test is a construct.

They go on to specify that the variable to be measured is constructed by a universe of content, from which we need to sample and design items that are domain-relevant, develop from these a test that is representative of the domain and provides an adequate sample of the domain. Psychometricians have focused their attention on whether these and other technical requirements of tests are met. Wiliam (2010) argued that we need to shift the debate from the technical features of assessment (how well we are assessing) to discussions about what we are assessing. Sharper conceptualisation of the construct of

interest might then produce better instruments, with higher validity. Test validity has been defined in countless ways (Newton & Shaw, 2014); however, for the purposes of this chapter, validity refers to ‘the extent to which a test measures what it purports to measure’ (Kelley, 1927).

The domain of interest for educational assessments is typically defined by a curriculum, a syllabus or sometimes more loosely by the discipline more generally. Some educational assessments are not curriculum-related so the domain is ill-defined in those cases, but in many instances there is little explicit definition of the domain and instead the textbooks and assessment materials come to define the domain. From the domain, a testing framework can be constructed which indicates how the domain is operationalised in the test. Three distinct approaches to developing constructs can be found in the literature, although oftentimes they are blended in assessment practice:

1. Theory-based: constructs are formulated on the basis of a theory, such as Piagetian notions of the development of scientific ideas
2. Empirically-driven: constructs are devised from the results of students on previous tests
3. Views of subject-matter experts: constructs represent disciplinary and/or assessment experts’ views on what counts as progression in the subject

One example of a development that blends all three approaches is described by Briggs and colleagues (2006). They outlined a cognitive theory-based assessment that was built upon Sadler (1998) Project STAR Astronomy Concept Inventory. Data were collected on students’ developments of understanding in astronomy between 8th and 12th grade and information was also collated about their misconceptions. Design principles from the Berkeley Evaluation and Assessment Research (BEAR) Center were then used to compile items founded upon the theoretical and empirical understandings. These design principles

included the engagement of disciplinary experts with empirical data on students' responses to items. Their resulting construct map indicates that at level 5, students can put the motions of the earth and moon into a complete description of the motion in the solar system, explaining the day and night cycle, the phases of the moon (including the illumination of the moon by the sun) and the seasons. At the lowest level, there is no evidence of understanding, or responses are not relevant, but at level 1 the student does not demonstrate understanding of the systematic appearance of objects in the sky and may not know that the earth is spherical. Some possible misconceptions are given, including that the sun goes below the earth at night.

This is an impressive and rigorous approach to building a construct which credits progress in knowledge of astronomy at this level in a tightly defined domain. Few assessments are constructed in this way, in part due to the level of resources required to do so. Indeed, given the widespread nature of testing, there is remarkably little reflection on the constructs underpinning them. Far more attention is given to the resulting numbers, which take on a life of their own (Hanson, 2000). One example of this, though many could be cited, is the use of PISA results in the report *PISA 2012 Results in Focus. What 15-year-olds know and what they can do with what they know* (OECD, 2014). The content of the report goes on to produce detailed analyses of scores and levels, but there is very little reference to what students know and what they can do with it.

OECD claims that PISA is not curriculum-based (Bybee & McCrae, 2011) and that it rather measures the extent to which fifteen-year olds are capable of applying knowledge to everyday situations as well as their readiness to become engaged citizens (OECD, 1999). Assessment development is guided by subject-specific frameworks which provide ample description about the subject-domain, the constructs measured, the areas of assessment, the

test structure and format and response guides (OECD, 2009b). The development of the PISA science 2006 assessment was informed by a scientific literacy framework developed by OECD based on the most recent trends in science education at the time. The framework and the constructs it defines are further elaborated in the next section of this chapter. These will be explored in terms of the extent to which they emphasise language as part of what is measured in the PISA assessments.

2.2.1 PISA's scientific literacy framework and constructs

PISA science assessments are developed based on a framework focusing on scientific literacy developed by science education experts forming the PISA 2006 Science Expert Group. The framework and the literature supporting it are detailed in OECD (2006) and Bybee et al. (2009). The following quotes how scientific literacy is defined in PISA.

Scientific literacy is an individual's:

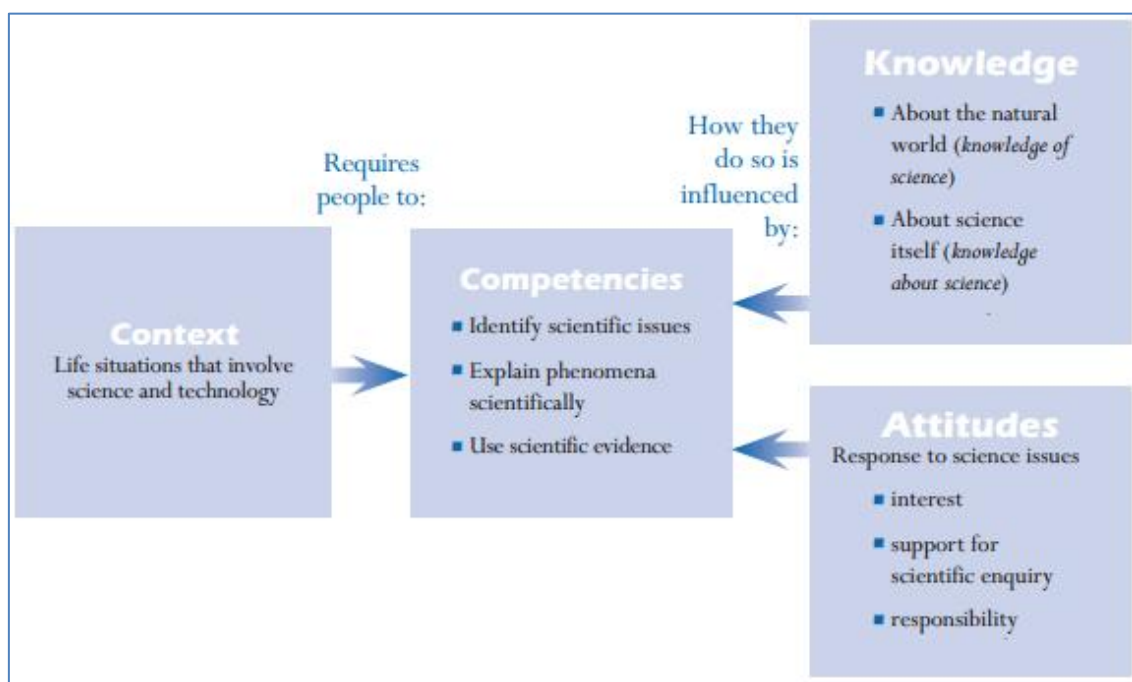
- Scientific knowledge and use of that knowledge to identify questions, acquire new knowledge, explain scientific phenomena and draw evidence-based conclusions about science-related issues,
- Understanding of the characteristic features of science as a form of human knowledge and enquiry,
- Awareness of how science and technology shape our material, intellectual, and cultural environment,
- Willingness to engage in science-related issues and with the ideas of science, as a reflective citizen.

(OECD, 2006, p.23)

Figure 2.1 below is a reproduction of the framework of scientific literacy used in PISA science 2006. PISA refers to constructs in its science assessments as scientific competencies. The scientific literacy framework used in the 2006 cycle is operationalised into the following three competencies: (1) Identifying scientific issues; (2) Explaining

phenomena scientifically; (3) Using scientific evidence (OECD, 2006). The competencies are associated with the practice of science and are related to essential abilities such as reasoning and critical thinking skills, transformation of data to tables and graphs, building arguments and explanations based on evidence, etc. (Bybee et al., 2009). They are directly related to science knowledge (scientific issues, phenomena) and science processing skills (identifying scientific issues, providing scientific explanations, using evidence).

Figure 2.1: Scientific literacy framework of PISA 2006 science assessment



Source: OECD, 2006, p.26

Figure 2.1 indicates that performance on these competencies is influenced by an individual's knowledge *of* science (that is, knowledge about the natural world) and the knowledge *about* science itself as well as his/her attitude towards science. This puts science knowledge and thinking skills at the centre of the framework. Figure 2.1 associates the context with the science and technology reality in which individuals are embedded in their daily life. It makes less explicit the influence of individuals' culture. Moreover, Figure 2.1 does not put forward the role of communication in science. Since language is

not explicitly included in the way the framework is formulated and operationalised and it does not appear in the way constructs are articulated then language is not intended to be part of the construct in PISA science assessments.

In this part, I briefly examined the scientific literacy framework adopted by PISA and how it has been operationalised into science constructs. I noted that the role of communication in science is undermined and that language does not seem to be part of the science construct; at least not an intended one. This conclusion could have been satisfactory had PISA assessments been less dense in language. Indeed, a remarkable feature of PISA assessments is the density of the language in the assessment and the high expectations of written answers to open-ended questions. These aspects result in great reading and writing demands which could be so high that they could impact the level of difficulty of the questions. In the next section, I investigate the density of the language in PISA assessments and contrast it with the language load in TIMSS Grade 8 (population 2) tests. TIMSS is a comparable international test conducted by IEA. The test is curriculum-based and measures maths and science achievement across countries and its Grade 8 tests are administered to fourteen-year-olds, a comparable age group to the one targeted in PISA (fifteen-year-olds). I also compare the expected length of open-ended responses by examining the scoring guides of example questions in both PISA and TIMSS science tests.

2.2.2 Reading and writing demands in PISA assessments

Examining the language load in PISA science units and items gives a good indication of the reading demands imposed when reading PISA science tests. Likewise, exploring examples of scoring guides and model responses provides insight into the expected writing skills and challenges imposed on students while answering questions. The reading demands of a released PISA item were explored using the Pearson Reading Maturity

Metric (RMM) (Landauer et al., 2011; Landauer, 2012). RMM uses computational language models to estimate a measure referred to as ‘Word Maturity’. Word Maturity depicts how much experience is needed to reach an adult’s understanding of words, sentences and texts. The Word Maturity metric is highly correlated with well-established vocabulary tests ($\rho \sim 0.8$) such as Kaufman Brief Intelligence Test – II² (Landauer et al., 2011). Results were compared to the Pearson RMM results of similar analyses carried out on a TIMSS (Grade 8) science item. The writing demands imposed by PISA and TIMSS science assessments were compared in terms of length and expected complexity. The next section describes the results of these comparisons.

2.2.2.a Reading demands of PISA items

PISA items are organised within units; units are sets of items grouped based on a common stimulus and scored independently (see a detailed description in Chapter 5). A stimulus often consists of a text passage as well as a photograph, a figure, a graph, a table, etc. Figure 2.2 below presents the stimulus of a released PISA science unit and one of its items.

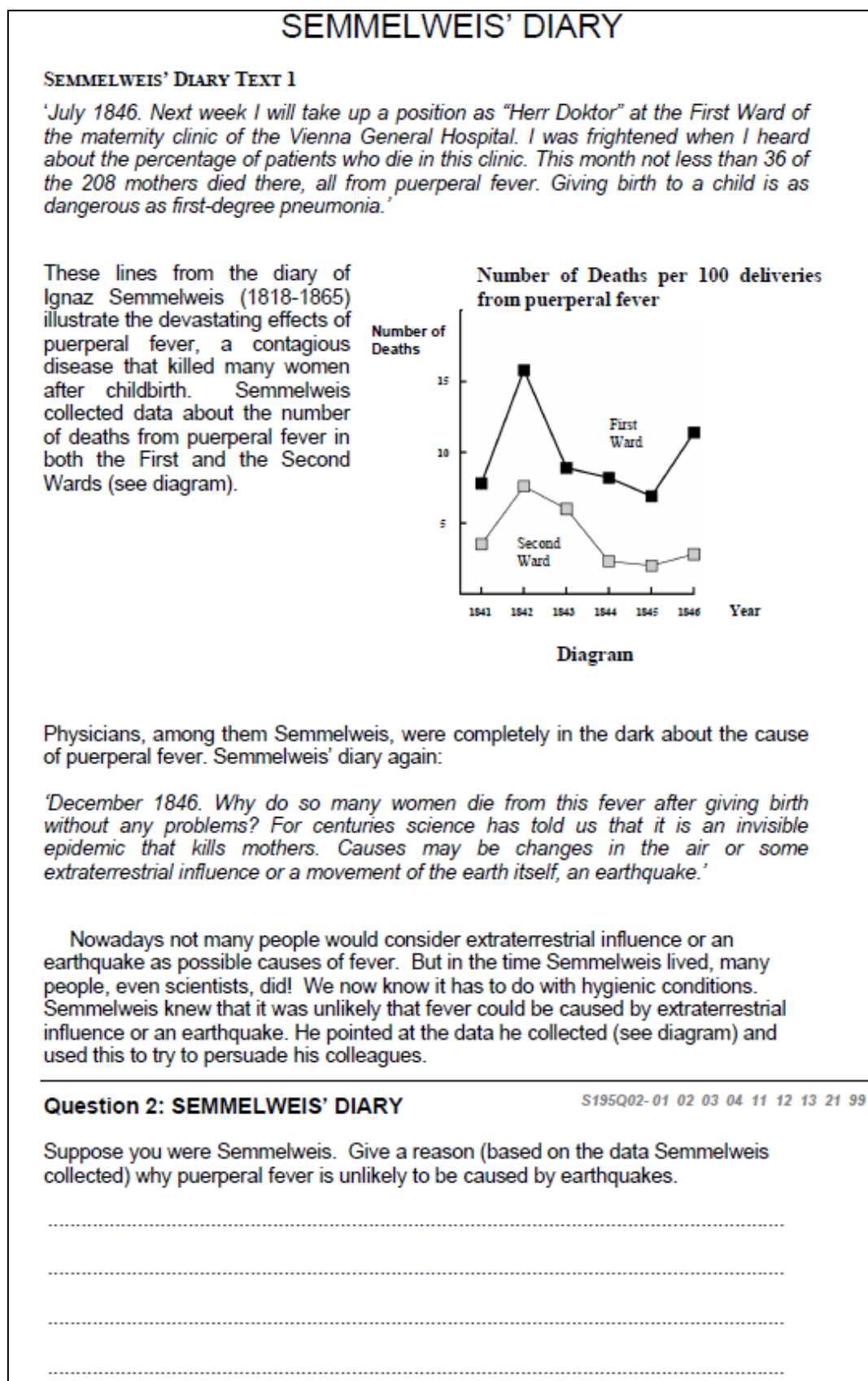
The stimulus consists of a text providing some context for a science issue, infectious diseases, and a graph representing scientific data, in this case number of deaths in two different wards as a function of years. With the exclusion of the graph and the title, the text consists of 259 words and 17 sentences and the question consists of 23 words and 2 sentences. A quick analysis of the readability level of the text using the Pearson Reading Maturity Metric (RMM) software³ revealed an overall score of 9.9 suggesting that the text is aimed at Grade 9 to 10 students, that is, age 15-16. In other words, an average 15-year old student should not find this text challenging. This corresponds to the age group of

² Kaufman, A. S. and Kaufman, N. L. (1990) Kaufman Brief Intelligence Test—Manual (American Guidance Service, Circle Pines, MN).

³ Software available online free: <http://www.readingmaturity.com/rmm-web/main#/passage/28156>.

students participating in PISA. However, to eliminate the possible effect of language on test performance, the text should be made simpler in a way to target a younger age group (e.g. 13-14 year olds) without leading to boredom. Technical words such as ‘puerperal’, ‘hygiene’ and ‘extraterrestrial’ were highlighted by RMM as being complex. The term ‘puerperal’ is found in the question and this might impede its understanding.

Figure 2.2: Released PISA science unit S195⁴



⁴ See the following website for released PISA items and scoring: www.oecd.org/pisa/38709385.pdf.

Items in TIMSS are not as dense. Figure 2.3 shows a science item from TIMSS 2003 administered to a comparable age group (Grade 8; that is, 14-year-olds) and covering the same topic as the PISA item (health, immune system).

Figure 2.3: TIMSS item S022154⁵ (Grade 8)

Transmission of cold in classroom

Scott went to school with a cold. Several days later, half of his classmates also had colds. What is one likely reason some classmates had colds but others did not?

The context is portrayed in the first two sentences and the question is one-sentence long and consists of 30 words. A quick readability test using Pearson RMM resulted in a score 3.1. This means that the text was targeted to Kindergarten to Year 1 (K-1), that is, students of age 5 to 6. The word ‘cold’ was highlighted as a relatively challenging word. This is possibly a challenging word for K-1 students but definitely not for 14- or 15- year olds. In TIMSS, language is simplified to a great extent to avoid it creating a construct-irrelevant variable. It can be argued that in TIMSS, language is dumbed down at the risk of not activating the necessary scientific domain but this aspect is beyond the scope of the current discussion.

When comparing PISA and TIMSS items, it is clear that solving PISA science items requires a solid level of language proficiency. Although the readability level does not exceed the average readability level of the participating age group, yet it is challenging enough to possibly affect performance on the science construct. Language demands are not restricted to reading in PISA; they are also reflected in scoring of items requiring extended

⁵See the following website for released TIMSS Grade 8 items and scoring (2003 cycle): <https://nces.ed.gov/TIMSS/educators.asp>.

responses. The following section compares the types of responses expected in PISA and those expected in TIMSS in terms of language demands.

2.2.2.b Writing demands in open-ended responses

Language demands in PISA do not only exist at the reading stage. Participants are required to produce answers that necessitate solid writing skills. A scoring guide is available for every PISA item; the guide provides a description of what constitutes a correct answer and what is considered incorrect.⁶ The scoring of items requiring short or extended answers provides criteria for what constitutes a full credit response, a partial credit response and what is considered as incorrect response with no credit. The scoring of the item discussed in the previous section (S195Q02 Semmelweis Diary) is presented below.

To get full credit (two marks), the PISA rubric indicated that students have to write something like *‘refers to the difference between the numbers of deaths per 100 deliveries in both wards, due to the fact that the first ward had a high rate of women dying compared to women in the second ward, obviously show that it had nothing to do with earthquakes. Not as many people died in ward 2 so an earthquake could not have occurred without causing the same number of deaths in each ward’*. Partial credit is given to answers such as *‘It would be unlikely to be caused by earthquakes because earthquakes don’t occur frequently’*. If students only state that earthquakes cannot cause the fever, they will not be rewarded any marks. In this item, in addition to the reading load, participants are also required to formulate longer responses and articulate clear cause and effect relationships. This ‘genre’ of writing in science is more demanding than for instance describing facts (Yore & Treagust, 2006).

⁶ See the following website for released PISA items and scoring: www.oecd.org/pisa/38709385.pdf.

In TIMSS, expected responses are shorter. For instance for item S022154 shown earlier, an example of correct response is: *Some students were hanging around Salil [sic] with him sneezing his germs onto them. The ones exposed to the virus caught it.* This item does not offer partial credit answers. Incorrect responses refer to something like *Some of his classmates did not like him so probably were not near him a lot.* Although this question also requires giving a cause, however, the expected correct response in TIMSS has a more anecdotal style and does not require including information from a graph. It is considerably shorter than the one expected in PISA (two sentences consisting of twenty-one words in TIMSS full credit response as opposed to two sentences comprising seventy-one words in the PISA full credit response).

The design of PISA items has been criticised for involving too much reading (Kjærnsli & Lie, 2004). Although most PISA science items are multiple choice questions, yet a third of them consists of open-ended questions (OECD, 2009b) and hence requires the type of responses described above. These reading and writing demands in science are not explicitly mentioned in the scientific literacy framework adopted by PISA nor are they referred to in the articulation of PISA science constructs. In other words, language is an unintended part of the construct in PISA science. However, the high reading and writing demands are likely to make language a construct-irrelevant variable in PISA items; that is, language is likely to affect the difficulty of the item. Nevertheless, when reviewing the literature on scientific literacy, the language demands in PISA no longer come as a surprise.

In the next section, I review the literature on scientific literacy and highlight the importance of language skills in acquiring scientific literacy. Given the importance of language in science education, I argue that language is part of the science construct

irrespective of the examiner's intention. I use the literature to reflect again on the PISA framework and competencies and show that language is implicitly included in the formulation of the PISA scientific framework and constructs. The prominent role of language is also detected in PISA proficiency levels.

2.3 A closer look at PISA's scientific literacy framework and proficiency levels

At first sight, PISA's scientific literacy framework and competencies do not seem to incorporate language skills. Reading, writing and communication of scientific knowledge are not explicitly spelled. In this section, I highlight the importance of language skills in science education by reviewing briefly the literature on scientific literacy and reflecting on the PISA framework and constructs. I support this argument further by examining the proficiency levels described by PISA and highlight how high performance in science is linked to acquiring good command of scientific language skills.

2.3.1 The fundamental sense of science literacy

Over the past couple of decades, the focus of science curricula has moved away from the traditional body of scientific knowledge to increasingly promote the concept of scientific literacy (Holbrook & Rannikmae, 2007). The rationale for that move was a belief amongst science education researchers that the science curriculum should aim to produce more scientifically literate citizens that are able to solve real-life problems and be engaged in science-related issues. Scientific literacy has quickly become the overarching framework for designing science curricula (Bybee, 1997; DeBoer, 2000). In England, Millar and Osborne's (1998) *Beyond 2000* report led to the implementation of a national science education curriculum in 2003, '*Twenty-First Century Science*', emphasising scientific

literacy for fifteen- and sixteen-year old students. Data gathered at the end of the first and second year of the curriculum implementation as part of a review of its effectiveness suggested that teachers endorsed the move towards a focus on scientific literacy (Millar, 2006). The research indicated that students showed more interest and engagement towards science.

Comprehensive definitions of science literacy have been suggested (cf. Bybee, 1997; DeBoer, 2000; Millar & Osborne, 1998; Norman, 1998). Below is the definition proposed by the National Research Council (1996, p.22)

Scientific literacy is defined in the *National Science Education Standards* as follows:

Scientific literacy means that a person can ask, find, or determine answers to questions derived from curiosity about everyday experiences. It means that such a person has the ability to describe, explain, and predict natural phenomena. Scientific literacy entails being able to read with understanding articles about science in the popular press and to engage in social conversation about the validity of the conclusions. Scientific literacy implies that a person can identify scientific issues underlying national and local decisions and express opinions that are scientifically and technologically informed. A literate citizen should be able to evaluate the quality of scientific information on the basis of its source and the methods used to generate it. Scientific literacy also implies the capacity to pose and evaluate arguments based on evidence and to apply conclusions from such arguments appropriately.

(NRC, 1996, p.22)

Most of the definitions of scientific literacy converge on the need to emphasise the interaction between science, technology and society (NRC, 1996; Thomas & Durant, 1987) and the importance of understanding the nature of science (Holbrook & Rannikmae, 2007), that is, science is empirical, tentative, value-laden, etc. (Lederman, 1992). They also agree on promoting science processing skills such as critical thinking, formulating hypotheses, drawing evidence-based conclusions (NRC, 1996; Thomas & Durant, 1987).

While most of these skills involve a form of communication of scientific knowledge, very few definitions of scientific literacy explicitly mentioned this aspect. In the following, I will examine the importance of reading and writing in acquiring the scientific literacy and later I describe the scientific language and describe its complexity based on research carried out on the readability of science textbooks.

2.3.1.a Reading and writing in science

Thomas and Durant (1987) included gaining a solid foundation in the scientific language as one of the eight criteria of acquiring scientific literacy. The importance of scientific literacy in its literal meaning; that is, reading and writing in science was also suggested by a small number of authors, such as Millar and Osborne (1998) and the American National Research Council (1996). This focus remained instrumental; limited to the ability to read science texts in newspapers and magazines or more functional making reading and writing as skills used by scientists (Anderson, 1999). Gee (2000) maintains that science imposes on learners the need to communicate using written or oral language as well as other symbol systems. Language is hence essential in science education.

Norris and Phillips (2003) considered that acquiring the scientific language and being able to communicate in science are essential to the development of scientific literacy. They argued that acquiring a solid grounding in reading and writing in science supports the consolidation of scientific knowledge and content as well as the development of science process skills. Norris and Phillips referred to the communication aspect of scientific literacy as being its *fundamental sense* and opposed it to the *derived sense* which included scientific knowledge, content and processes. They claimed that the fundamental sense reinforced the derived sense and vice versa; this relationship was described as symbiotic by Gee (2000). Engaging with oral and written arguments, researching information from

written material (print or digital), writing reports, notes and articles and communication of data in the form of tables, graphs, drawings, diagrams are at the heart of what scientists do and how students learn science (Anderson, 1999; Yore & Treagust, 2006).

Norris and Phillips (2003) contended that reading and writing in science are not simply functional skills that allow scientists to do science; reading and writing are *constitutive* of science. In other words, science endeavour is not possible without them. They maintained that reading does not consist of mere ‘skilled word recognition’ (p.227). Rather, reading consists of ‘comprehending, interpreting, analysing, and critiquing texts’ (p.229) many of these mental activities being central to science (Gaskins et al., 1994a, b). Hence, the relationship between science and reading is very intimate (Norris & Phillips, 2003). Writing is what puts thoughts in print so that their coherence can be examined, reflected on, debated and interpreted (Halliday & Martin, 1993; Olson, 1996). Norris and Phillips (2003) argue that writing is what makes these activities possible.

2.3.1.b The complexity of the scientific language

Research carried out with native speakers of English on the readability of science texts identified various lexical and grammatical features characterising the scientific discourse and challenging students’ understanding of science written material (Halliday & Martin, 1993; Chavkin, 1997; Fang, 2006; Groves, 1995; Labasse, 1999; Merzyn, 1987; Sutton, 1998; Swaffar, 1988). There is no consensus on whether lexical features or syntactic patterns are chiefly responsible for any ambiguity in science text readability. According to Halliday and Martin (1993), lexical difficulties are overrated as the new complex vocabulary is overstressed in classrooms and is not in itself problematic. It is rather the relationships established among the lexical items as conveyed by grammar that render them more complicated. However, many studies challenged Halliday and Martin’s (1993)

stance by providing empirical evidence that students are challenged by the lexical components of science texts (Chavkin, 1997; El Masri & Vlaardingerbroek, 2010; Fang, 2006; Ginsguger-Vogel & Astolfi, 1987; Groves, 1995; Merzyn, 1987; Sutton, 1998).

Halliday and Martin (1993) outlined in detail the various grammatical features peculiar to the scientific discourse that seem to affect the clarity of textual material. These include interlocking definitions, technical taxonomies, special expressions, lexical density, syntactic ambiguity, grammatical metaphors and semantic discontinuity. Moreover, Fang's (2006) extensive work with lower secondary school students identified additional grammatical features peculiar to science discourse that appear to hamper students' readability of texts. Prepositions, conjunctions and pronouns often imply causal, consequential or chronological relationships which are ambiguous for students to infer when reading.

The use of metaphors, ellipses and nominalisations, which help scientists pack information, convey precision and build generalisations, result in complex sentences that are very dense in information, with long noun phrases and multiple subordinate clauses (Chavkin, 1997; Fang, 2006; Groves, 1995) making scientific knowledge abstract for students. In addition, Fang (2006), Moss (2000) and Sutton (1998) pointed out that the passive voice used in scientific texts in order to ensure the objectivity of information seems to alienate readers. Although all these features were identified in studies investigating the readability of science textbooks, they are all intrinsic to the scientific discourse which is critical for the acquisition of scientific literacy (DeBoer, 2000; Norris & Phillips, 2003). Hence, these features are expected to permeate into the language used in the formulation of examination questions, such as PISA assessments, making them less accessible to candidates.

Groves' (1995) analysis of narrative texts in four frequently used science textbooks suggested that science texts are heavily loaded with new vocabulary. Moreover, Chavkin (1997), Fang (2006) and Sutton (1998) noted that common words have very specific definitions in physical sciences yet they are used differently in everyday contexts. For instance, words such as 'force', 'energy' and 'weight' have technical meanings in the physical sciences but are used loosely in everyday language. When reading science texts involving such words, common language often interferes with the scientific meaning and creates confusion in readers' minds. Readers must then identify the context in which these terms are used (scientific or everyday language), then avoid the connotations inherent in the everyday meaning of the term. Since the heavy load of vocabulary and the common ambiguous terms are part of scientific language, it is reasonable to suppose that they would also appear in examination questions such as PISA science assessments.

Similar studies evaluating the readability of science textbooks written in French with lower secondary French native speakers determined comparable problems. Ginsguger-Vogel and Astolfi (1987) reported that science texts written in French were loaded with new vocabulary and technical as well as non-technical terms. Besides, students in this study appeared to be discouraged by sentences that contained a lot of information as they found it difficult to select the relevant clues and establish relationships. Thus, it is likely that students reading science examination questions in French may encounter similar challenges compared with their English-speaking counterparts. No similar studies investigating the readability of science textbooks in Arabic have been published to this date. However, Halliday and Martin (1993) showed that similar features of the scientific language exist in other languages such as Chinese. It is very likely that science texts written in Arabic will include syntactic and lexical features that convey precision and objectivity of knowledge.

Based on the discussion of how constructs are built in the first section of this chapter, it can be said that the centrality of language in science education cannot but be infiltrated in constructs designed based on a science literacy framework. Language is part of science constructs, shaped around science literacy irrespective of the intention of the examiner. This conclusion can also be extended to PISA science tests. The following section discusses the language aspect in PISA scientific literacy framework and constructs.

2.3.2 The language element in PISA's scientific literacy framework

PISA defines scientific literacy as follows:

Scientific literacy is the capacity to use scientific knowledge, to identify questions and to draw evidence-based conclusions in order to understand and help make decisions about the natural world and the changes made to it through human activity.

(OECD, 1999, p.60)

Although not plainly stated in PISA's definition of scientific literacy, it would be hard to think about how individuals would acquire knowledge, explain phenomena scientifically, draw conclusions, engage with science related issues and reflect on them with poor language skills. The PISA competencies are further broken down into features as shown in Table 2.1 below.

Table 2.1: PISA 2006 scientific competencies

Identifying scientific issues	Explaining phenomena scientifically	Using scientific evidence
Recognising issues that is possible to investigate scientifically	Applying knowledge of science in a given situation	Interpreting scientific evidence and making and communicating conclusions
Identifying keywords to search for scientific information	Describing or interpreting phenomena scientifically and predicting changes	Identifying the assumptions, evidence and reasoning behind conclusions
Recognising the key features of a scientific investigation	Identifying appropriate descriptions, explanations, and predictions	Reflecting on the societal implication of science and technological developments

Source: OECD, 2006, p.29

In keeping with Norris and Phillips' (2003) distinction between fundamental and derived sense of scientific literacy, some of the above aspects of the constructs are directly related to reading, speaking or writing in science. For instance, '*Identifying keywords to search for scientific information*' is directly related to reading skills in science. Also, '*Describing or interpreting phenomena scientifically and predicting changes*' requires writing skills that express clear descriptions and interpretations. Note that descriptive texts are a different genre from argumentative texts expressing interpretations based on data. Moreover, '*Interpreting scientific evidence and making and communicating conclusions*' requires great reading proficiency; being able to read scientific data in different forms (text, graphs, tables). It also requires communication skills to express conclusions verbally or in form of a coherent text.

Hence, although PISA does not include a construct that clearly states that communication skills are measured, the fundamental sense of scientific literacy is weaved into the constructs that relate more overtly to the derived sense. Language is hence not a mere construct-irrelevant variable in PISA science tests as discussed in Section 2.2.2; it is an

intrinsic part of the science constructs. The centrality of language skills is also reflected in the proficiency levels of scientific literacy described in PISA manuals. The next section sheds light on this aspect.

2.3.3 Language expectations in proficiency levels of PISA science

Constructs reflect the domain of interest and indicate progress in the domain. As such, constructs are the basis upon which scoring criteria can be developed. Grading criteria or level descriptors are effectively depictions of the construct. In PISA science, scores represent six proficiency levels of scientific literacy, which include being able to explain, justify and develop arguments. Figure 2.4 below presents PISA science proficiency levels.

In Figure 2.4, I have put in bold terms and phrases that refer to skills requiring reading, writing and verbal communication in science; that is, elements of the fundamental sense of scientific literacy. It can be noted that the six levels make direct reference to the fundamental meaning of scientific literacy. Moreover, one can notice that increasing performance in science requires higher proficiency in scientific language. Let us look at the progression by examining three levels; the lowest, the highest and one representing medium proficiency.

Proficiency Level 1 consists of straightforward presentation of simple scientific explanations. This level requires basic reading skills of simple scientific texts and the ability to communicate what has been read clearly. Level 3 involves slightly more advanced reading skills which enable examinees to recognise scientific issues when reading simple descriptions of them. They need to select facts and formulate explanations based on the facts. These skills are more advanced than simple presentation of scientific explanation provided by a text. The highest proficiency level (Level 6) entails very

advanced reading and writing ability skills that allow examinees to identify scientific knowledge as well as to integrate scientific information from different sources. This level requires examinees to express thorough scientific logic and explain scientific ideas as well as articulate clear convincing arguments. Based on the three selected proficiency levels, one can conclude that the progression of the fundamental sense of scientific literacy mirrors that of the derived sense in PISA proficiency levels.

Figure 2.4: PISA proficiency levels in scientific literacy

At Level 6, students can consistently **identify, explain** and apply scientific knowledge of and knowledge about science in a variety of complex life situations. They can **link different information sources and explanations** and use evidence from those sources to justify decisions. They clearly and consistently **demonstrate advanced scientific thinking and reasoning**, and they use their scientific **understanding** in support of **solutions to unfamiliar** scientific and technological situations. Students at this level can use scientific knowledge and **develop arguments** in support of recommendations and decisions that centre on personal, social or global situations.

At Level 5, students can **identify** the scientific components of many complex life situations, apply both scientific concepts and knowledge about science to these situations, and can **compare, select** and **evaluate** appropriate scientific evidence for responding to life situations. Students at this level can use well-developed inquiry abilities, **link** knowledge appropriately and **bring critical insights** to situations. They can **construct explanations based on evidence and arguments based on their critical analysis**.

At Level 4, students can work effectively with situations and issues that may involve explicit phenomena requiring them to **make inferences** about the role of science or technology. They can **select and integrate explanations** from different disciplines of science or technology and **link** those explanations directly to aspects of life situations. Students at this level can **reflect** on their actions and they can **communicate** decisions using scientific knowledge and evidence.

At Level 3, students can **identify clearly described** scientific issues in a range of contexts. They can **select** facts and knowledge to **explain** phenomena and apply simple models or inquiry strategies. Students at this level can **interpret** and use scientific concepts from different disciplines and can apply them directly. They **can develop short statements** using facts and **make decisions** based on scientific knowledge.

At Level 2, students have adequate scientific knowledge to **provide possible explanations** in familiar contexts or **draw conclusions** based on simple investigations. They are capable of **direct reasoning** and **making literal interpretations** of the results of scientific inquiry or technological problem solving.

At Level 1, students have such a limited scientific knowledge that it can only be applied to a few, familiar situations. They can **present** scientific explanations that are obvious and follow explicitly from given evidence.

Source: OECD, 2013a, p.113

Having the fundamental sense of language clearly stated in the proficiency levels in science supports the argument that language is a construct-relevant variable in PISA science assessments. This conclusion has significant implications on the way PISA results are interpreted. Language is assumed to be tangential to science constructs and neutral in terms of its impact on the demands and difficulty of an item. This assumption becomes more problematic when considering the differential impact of language on the level of

difficulty and demand of a question. The next section examines the relationship between language and thinking based on the linguistic relativity hypothesis. This hypothesis suggests that different languages impact thought in different ways and confer different cognitive demands. This hypothesis is illustrated using two examples of science test items.

2.4 Are science constructs comparable across languages?

The presence of a common underlying structure amongst all languages has driven much of the research in cognitive science and learning. These shared properties have been referred to as ‘language universals’ (Chomsky, 1986; Pinker & Bloom, 1990; Pinker, 1994). Language universals have been however criticised by various researchers. Evans and Levinson (2008) provided a comprehensive review of the literature and contended that claims for language universals do not stand on accurate empirical evidence. Indeed, in their title, they referred to this theory as a ‘myth’. They argued that languages are fundamentally distinct on every level (sound, grammar, lexicon, meaning) and that the features they shared are only a few and superficial.

This area of research is of particular relevance to my research as it frames many learning theories to which educational assessment is directly associated. Evans and Levinson (2008) maintain that cognitive science would benefit tremendously if the concept of language diversity would be embraced. In international testing, language diversity makes the comparability of language versions of the same test even more challenging. Evans and Levinson’s (2008) argument drew on laws of population biology. They contended that the diversity in language reflects the immense variability in culture as language is a cultural element. They maintained that the ‘bio-cultural hybrid’ (p.446) feature of language is its most important property.

The diversity of language and its tight relationship with culture has been hypothesised by Wilhelm von Humboldt in the nineteenth century. This hypothesis is referred to as the linguistic relativity principle, or more commonly the Sapir-Whorf hypothesis, or Whorfianism. This principle contends that the structure of a language influences the world view of its speakers and their cognitive processes. There are various versions of this principle. On one end the strong version, also referred to as ‘linguistic determinism’ holds that language determines thought and constrains thinking (Carroll, 1956). This version is attributed to Benjamin Lee Whorf and Ludwig Wittgenstein and can be well detected in one of Wittgenstein’s (1922) famous quotes ‘*The limits of my language mean the limits of my world*’.

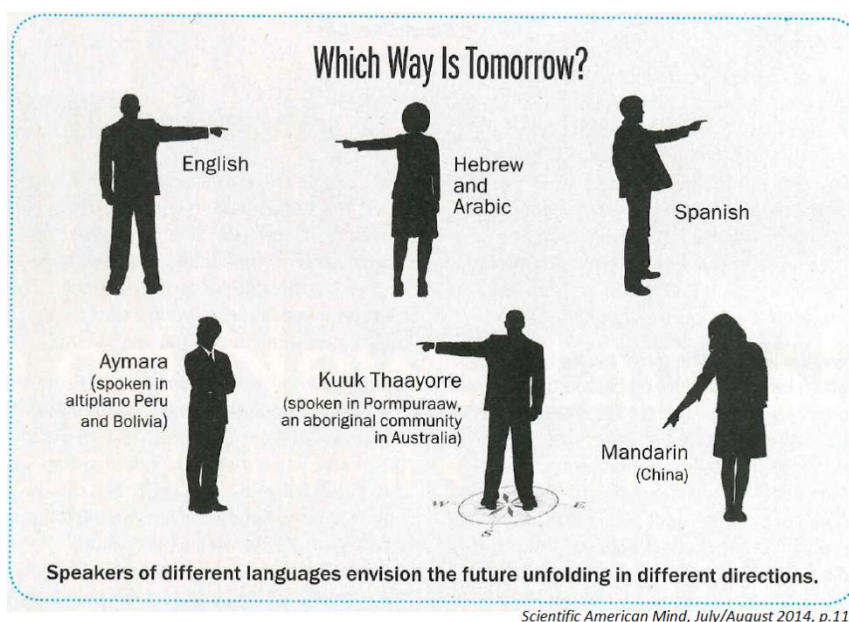
Linguistic determinism has been criticised widely because of poor or contested empirical evidence. A weaker version of this hypothesis seems more appealing; it considers linguistic categories and usage as only shaping thoughts but not determining them. Empirical studies seem to be more in favour of the watered down version of linguistic relativism. In this thesis, I adhere to the weak version of linguistic relativity; that is, different languages, having different syntactic and lexical features, influence thinking in different ways. In this section, I use the example of time perception across different cultures to illustrate linguistic relativism. Later, I provide examples of science items to relate this theory to assessment across languages.

2.4.1 Perception of time and linguistic relativism

One domain that has been quite successful in illustrating linguistic relativity is the domain of time. Humans tend to perceive time as a one-dimensional and directional entity that they conceptualise in terms of spatial references (Clark, 1973; Lehrer, 1990; Traugott, 1978). This aspect tends to be universal across cultures (Boroditsky, 2001). However, different

cultures talk about time differently. In English, people look *forward* to a meeting and put the past *behind* them. This suggests that time is perceived by English-speaking people as being on a horizontal axis where the past is at the *back* and the future is at the *front*. The back and front metaphor is also used in mandarin when talking about time. However, mandarin-speakers also use other spatial references. They use the mandarin equivalent of *up* to refer to express earlier events, and the mandarin equivalent of *down* to talk about later events (Boroditsky, 2001). Hence, mandarin-speakers tend to perceive time on a vertical axis. More examples of how different cultures perceive time are discussed in the literature. These are depicted in Figure 2.5 below taken from the *Scientific American Mind* (July/August 2014, p.11) showing how different cultures point to different directions when asked to point to the future: *Which way is tomorrow?*

Figure 2.5: Which way is tomorrow in different cultures



Source: *Scientific American Mind* (2014, p.11)

According to this article, English-speaking people tend to point to the left; Mandarin-speaking people point downwards. Spanish-speaking persons point to the front and Aymara-speaking persons look to the right. Arabic- and Hebrew-speaking individuals

point to the right; an interesting finding given that both languages are amongst the few that are written from right to left. Another interesting example is how an aboriginal community in Australia speaking Pormpuraaw language refers to time using cardinal directions (North, East, West and South). This mirrors how this community describes object locations and uses the same references when talking about time (Boroditsky & Gaby, 2010).

In summary, these examples suggest that ways of speaking or writing about time influence how humans perceive this concept. This hypothesis has great implications in educational assessment. If language influences thought and cognitive processes, this suggests that languages impact thinking in different ways and present different linguistic and cognitive challenges. This has direct repercussion on assessment across countries as different language versions of the same assessment might inherently present examinees from distinct language backgrounds with varying linguistic and cognitive difficulties.

2.4.2 Different languages: Different cognitive and linguistic demands

In assessment, linguistic relativism can be often seen in various cases, for instance when home language influences performance on some items. Luykx et al. (2007) reported findings from a study where Spanish-speaking students in USA were confused by the abbreviations for degrees Fahrenheit (F) and degrees Celsius (C). In this case, students would give answers such as 20F was colder than 20C. It appears that the confusion was because F related to the term *frío* in Spanish [cold] and C was associated with *caliente* [hot]. A similar ambiguity was depicted with Haitian students, but this time it was because of the French terms [*froid*] for cold and [*chaud*] for hot. The initials ‘F’ and ‘C’ are often present on bathroom faucets in Haiti to distinguish handles controlling the flow of cold and hot water.

This empirical evidence suggests that language affects the difficulty of questions and may introduce challenges that do not exist in other languages. In this case, speakers of Latin-based languages seem to be misled by the terms referring to temperature and appear to associate them with units of temperature which are unrelated to the terms themselves. This confusion has no basis in other languages such as Arabic for instance. The unit ‘degrees Celsius’ is translated into the equivalent of ‘degrees centigrade’; that is, ‘درجة مئوية’ [daraja mi’awiya] symbolised by the Latin letter ‘C’. The unit ‘degrees Fahrenheit’ is translated into ‘درجة فهرنهايت’ [darajat fahrenheit] symbolised by the Latin letter ‘F’. These units in Arabic cannot be confused with the terms ‘بارد’ [bared] for cold and ‘ساخن’ [sakhen] for hot because they do not start with the letters C and F respectively, let alone the fact that Arabic uses a completely different script to write these words.

The previous example highlights how certain language backgrounds can result in unexpected misunderstandings of assessment questions. These confusions do not result from poor translation or adaptation of the tests; they are rather a reflection of how different languages influence individuals’ understandings. Next, I provide another example of questions relating to the moon cycle which illustrates how languages impose different demands related to particular concepts. In this example, different cognitive and linguistic demands result from varying levels of word frequency in different languages. Low frequency words constitute an impediment for understanding questions (Graesser et al., 2006, Lenzner et al., 2010). Other problematic features will be discussed in Chapter 8. In the example below (see Figure 2.6a and b), I show how terms designating moon phases in three different languages – English, French and Arabic – impose different linguistic and cognitive demands on examinees in test items because of their differential frequency in the respective languages.

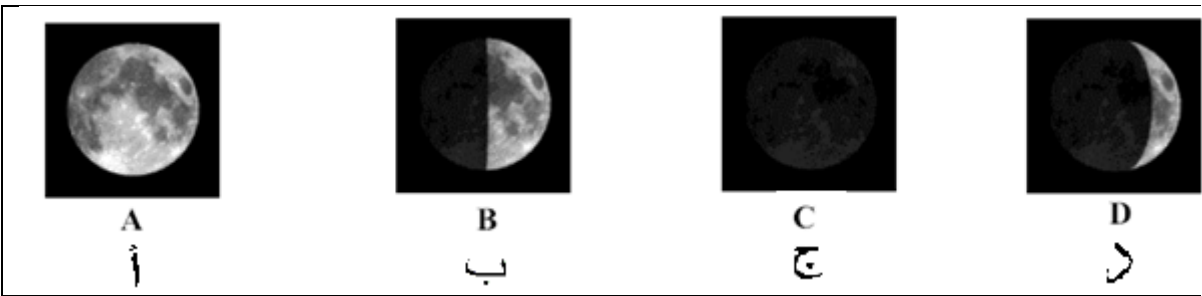
Figure 2.6a summarises a hypothetical lesson focusing on the lunar cycle typically covered between upper elementary to lower secondary schools. The moon phases are depicted in three languages English, French and Arabic.

Figure 2.6a: Moon Phases in different languages

									
English	<i>new moon</i>	<i>waxing crescent</i>	<i>first quarter</i>	<i>waxing gibbous</i>	<i>full moon</i>	<i>waning gibbous</i>	<i>last quarter</i>	<i>waning crescent</i>	<i>new moon</i>
French	<i>nouvelle lune</i>	<i>premier croissant</i>	<i>premier quartier</i>	<i>lune croissante gibbeuse</i>	<i>pleine lune</i>	<i>lune décroissante gibbeuse</i>	<i>demier quartier</i>	<i>demier croissant</i>	<i>nouvelle lune</i>
									
Arabic	المحاق	الهلال المتناقص	التربيع الثاني	الأحدب المتناقص	البدر	الأحدب المتزايد	التربيع الأول	الهلال المتزايد	المحاق

Let us consider a type of question that examinees at this age might face. Figure 2.6b includes three questions in English and their translation in French and then Arabic.

Figure 2.6b: Same questions in three languages



English	1) Which one of the photos above represents the <i>new moon</i> ? ____
	2) Which one of the photos above represents the <i>full moon</i> ? ____
	3) Which one of the photos above represents the <i>waxing crescent</i> ? ____
French	1) Parmi les photos ci-dessus, laquelle représente la <i>nouvelle lune</i> ? ____
	2) Parmi les photos ci-dessus, laquelle représente la <i>pleine lune</i> ? ____
	3) Parmi les photos ci-dessus, laquelle représente le <i>premier croissant</i> ? ____
Arabic	١- أي من الصور أعلاه يمثل <i>المحاق</i> ؟ [almahaq] ____
	٢- أي من الصور أعلاه يمثل <i>البدر</i> ؟ [albadr] ____
	٣- أي من الصور أعلاه يمثل <i>الهلال المتزايد</i> ؟ [alhila almutazayed] ____

In the questions above, examinees are instructed to fill in the blanks with the letter corresponding to the correct lunar phase. Let us now analyse the linguistic demands for each question in each language then compare results across languages. Within each language, the stem of the question is constant; that is, in English for instance the statement ‘Which one of the photos above represents the’ is the same in the three questions. This is the case in French and Arabic. Hence, my hypothesis is that the difference in the level of demands between questions within each language is determined by the demands imposed by the frequency of words in the lunar phase.

- English

In questions 1 and 2, the terms ‘new’, ‘full’ and ‘moon’ are high frequency words in English. Hence, the first two questions are easy to process and examinees are very likely to choose letter C for item 1 (moon with no light) and letter A for item 2 (moon ‘full’ of light). Examinees might hesitate between letters C and D for new moon but the term ‘new’ will allow them to eliminate A and B as possibilities.

Question 3 includes two less familiar terms. The word ‘waxing’ is associated with ‘adding wax’ and the word ‘crescent’ is not a common shape or word like triangle and square. So this question in English is likely to impose greater linguistic demands than the other two

phases of the moon. Examinees might not be able to easily eliminate choices especially if this question was not asked with the previous two.

- French

Similarly to English, the terms '*nouvelle*', '*pleine*' and '*lune*' are high frequency words in French and are not likely to impose language demands on examinees who would very likely be able to answer questions 1 and 2. Unlike in English, question 3 has also high frequency words. The term '*premier*' (meaning first in English) is very common in French and the term '*croissant*' relates to the famous French pastry which examinees at that age would easily recognise. Hence, in French none of the questions seem particularly demanding or imposing linguistic or cognitive challenges.

- Arabic

It is always useful to note in this discussion that Arabic is a diglossic language; i.e. the spoken language '*العامية*' [*ala'mmiya*] is quite different from the literary one, referred to as '*الفصحى*' [*al fuss-ha*] or the modern standard Arabic. When the language of instruction is Arabic in science, it is the standard Arabic that is adopted and not what students speak at home. The same is true for assessments; they are administered in the standard Arabic language. Hence, the frequency of words is also related to how similar the word in standard Arabic is to the one used in the colloquial language.

In question 1, the term '*المحاق*' [*almahaq*] has a very low frequency of usage in Arabic, especially the spoken one, and is only likely to be encountered in a context discussing the phases of the lunar cycle. Hence, examinees would not really have a clue about which phase this term would refer to. Question 2 includes the term '*البدر*' [*albadr*]. This term is quite familiar in Arabic cultures and the word is the same in standard and spoken Arabic. Often, the beauty of a woman is compared to the full moon to imply the brightness and

splendour of her face. Hence, question 2 should not entail any particular linguistic demands in Arabic. In question 3, the terms ‘الهلال’ [*alhilal*] and ‘المتزايد’ [*almoutazyed*] are quite high in frequency. The term [*alhilal*] is the same in standard and spoken Arabic. It means ‘crescent’ which is quite of a familiar shape in Arabic-speaking countries since it is the symbol of Islam and Arabic-speaking students are likely to be in countries with a majority of Muslims. The term [*almoutazayed*] is similar in standard and spoken Arabic. It means increasing and has also a high frequency in the language. Hence, in Arabic, question 3 does not impose any particular linguistic and cognitive demands.

To summarise the findings of the previous discussion, one can note that item 3 was the most linguistically demanding in English, item 1 was the most demanding in Arabic while none of the items seem to be particularly linguistically demanding in French. Within the same language, the level of demands of a question can vary with the level of frequency of a particular word referring to a given concept (Graesser et al., 2006; Lenzner et al., 2010). Let's assume that the stem of the question confers the same level of demands across languages; that is, the phrase ‘Which one of the photos above represents the’ and its translation into French and Arabic have no additional impact on the language demands. We can conclude that questions across languages may impose different linguistic and cognitive demands even if the translation has no impact. In this case, the translation of the terms from English to French and Arabic leads to different word frequencies in the target languages making some questions inherently less demanding and others intrinsically more challenging linguistically. Furthermore, the words in the question are specific technical terms with scientific meanings that cannot easily be exchanged for lower or higher frequency terms.

Obviously, the previous analysis is theoretical and it would be interesting to collect empirical evidence to see whether these language demands lead to differential item difficulties. The distinction between ‘demand’ and ‘difficulty’ here is one that Pollitt et al. (2007) make. Following Pollitt et al., by demands, I mean the cognitive and mental processes that examinees experience when carrying out a given task like answering a question. By difficulty, I mean the statistical measure given empirically by the proportion of successful students on a given task.

CONCLUSION

Educational constructs are commonly developed based on curricula or frameworks encompassing the most recent research in a particular discipline (Cronbach & Meehl, 1955). PISA science assessments are built based on a framework focussing on scientific literacy in line with the vast movement of science education towards less strictly content-focused curricula and greater promotion of science thinking skills (OECD, 1999). Constructs in PISA follow from the scientific literacy framework. Recent research in science education has raised awareness on the importance of language in scientific literacy. Reading, writing and speaking are constitutive parts of the scientific endeavour and cannot be separated from scientific literacy (Anderson, 1999; Gee, 2000; Norris & Phillips, 2003; Yore & Treagust, 2006). The importance of language is hence likely to be manifested in assessments designed using scientific literacy frameworks. Thus, a question worth asking is whether PISA considers language as part of the science construct or whether it is a construct-irrelevant variable.

A first glance at the scientific literacy framework and the science constructs as described by PISA imply that language is not part of the construct. However, examining the reading and writing demands of PISA in comparison with a similar international test, TIMSS

Grade 8, suggests that language is an important aspect of PISA. Moreover, exploring the language element in the way PISA formulates its scientific literacy framework and the science constructs in light of the literature on scientific literacy implies that language is an intrinsic part of science assessments, such as PISA. This argument is supported by the description of the proficiency levels in PISA which associate high proficiency in scientific literacy to high abilities in reading and writing in science.

If language is part of the science constructs in PISA, then having equivalent constructs across countries and languages becomes more challenging. Theories of language diversity (Evans & Levinson, 2008) which reject the presence of language universals hypothesised by several scholars (e.g. Chomsky, 1986; Pinker & Bloom, 1990; Pinker, 1994) make cross-language and cross-cultural comparisons very difficult. Besides, linguistic relativism suggests that different languages with their idiosyncrasies influence how individuals make meaning of the world (Carroll, 1956). Even a weaker version of this theory has grand implications on international assessments.

Despite rigorous translation and adaptation techniques, languages impose different linguistic and cognitive demands due to their own peculiar features and these differences may lead to differential item difficulties. Indeed, empirical studies on PISA and TIMSS detected different functioning of items across countries (Ercikan & Koh, 2005; Grisay & Monseur, 2007; Grisay et al., 2007; Hauger & Sireci, 2008; Huang et al., 2014; Kreiner & Christensen, 2014; Le, 2009; Oliveri & Ercikan, 2011; Sandilands et al., 2013; Wu & Ercikan, 2006; Xie & Wilson, 2008).

In this chapter, I used two examples to illustrate how different languages can impact the demands of an item when quality of the translation is not a problem. The next chapter discusses translation and adaptation of tests and points out that even with thorough

translation, comparability of constructs is very difficult to achieve. The challenges arise because of the various sources of bias that are inherent in the cultures and the adaptation techniques employed.

CHAPTER 3

Equivalence of Language Versions

‘Le diable bat sa femme et marie sa fille.’

‘The devil beats his wife and marries his daughter.’

(Literal translation from French)

INTRODUCTION

Scores from international large-scale assessments – such as PISA or TIMSS – are directly compared across countries. Following every cycle of PISA and TIMSS, league tables are produced in which countries are ranked based on the achievement of selected national samples on the overall tests and on every subject. This direct cross-country comparison is rooted in the assumption that the tests are strictly equivalent. However, criteria for establishing equivalence vary across theories and the degree of comparability can be set at different levels.

Under classical test theory (CTT), parallel forms of a test are considered to be the most equivalent forms of tests fulfilling strict criteria of comparability and allowing scores to be interchangeable across forms. Typically, parallel forms are administered to samples within the same population and hence are normally administered in the same language. They have not been developed in the context of cross-cultural assessments. Hence, justifying the direct comparison of scores across countries in the context of international assessment by considering the tests to be parallel is not a straightforward matter.

Revisiting the definition of parallel forms under CTT seems necessary to investigate whether the conditions allowing direct score comparisons across versions could be satisfied in the context of international large-scale assessments. In other words, it would be worthwhile examining whether language versions of the same international test could be

considered as parallel forms. CTT also includes more flexible variations of the parallel forms model: the tau-equivalent model, the essentially tau-equivalent model and the congeneric model. However, in the latter models, scores are not directly comparable and conclusions should be restricted to the degree of equivalence set when the tests were developed.

Under latent test theory (LTT), equivalent forms of the same assessment are produced by equating or linking assessments; that is, by mapping scores generated by the various forms of the same test on the same scale. Technically speaking, any two assessments can be equated or linked. The issue lies for some scholars in the strength of the link made between the tests while for others the issue relates to the nature of the link and its meaningfulness. Here too, interpretations of scores are valid to the extent allowed by the degree of equivalence amongst the various assessments.

I sign up to the second school of thought where equating or linking tests only makes sense when the construct measured is the same and the populations compared are similar. In Chapter 2, I argued that language is an intrinsic part of the science construct in science assessments including PISA, irrespective of the intention of the question setter. In this Chapter, I argue that because language is part of the construct, equating or linking language versions of the same test is akin to comparing tests with very different constructs. It is questionable how meaningful it is to equate or link different populations on tests covering distinct constructs.

The above discussion is conceptual because national versions of PISA tests are only equated across administration cycles and not across language versions. Indeed, national versions of international large-scale assessments are translated and adapted from at least one source version of the test. Sophisticated and rigorous methods of test translation and

adaptation are adopted to produce equivalent forms of the same test. In PISA for instance, a thorough double translation of source versions developed in English and French into other languages is employed (OECD, 2009a). Despite the diligence invested in the translation and adaptation methodology employed, researchers have reported issues with the translation of PISA tests and the occurrence of differential item functioning (DIF) in many cases (Grisay & Monseur, 2007; Grisay et al., 2007; Huang et al., 2014; Kreiner & Christensen, 2014; Le, 2009; Oliveri & Ercikan, 2011; Xie & Wilson, 2008). Similarly, issues associated with the translation and adaptation of tests were also revealed in several empirical studies on TIMSS tests (Ercikan & Koh, 2005; Hauger & Sireci, 2008; Wu & Ercikan, 2006) and PIRLS (Sandilands et al., 2013) as well as in national assessments in the UK (Ercikan et al., 2004; Jones, 1998).

The abundance of empirical evidence pointing to the poor level of equivalence across test versions administered in different countries is often overlooked and conclusions drawn based on these comparisons are increasingly influential at policy level. Educational systems in some countries such as Norway, Canada, Germany and Japan underwent major educational reforms following the publication of PISA results (Breakspear, 2012; Ertl, 2006; Schleicher, 2009; Westbury, 1992). For instance, the PISA shock in Germany following the publication of PISA 2000 results led to the introduction of national educational standards and quality assurance measures (Ertl, 2006). In Japan, the format of national tests was modified to include more open-response tasks similar to those in PISA (Schleicher, 2009).

In this chapter, I review the literature on parallel tests and argue that the standard definition of parallel forms cannot be extended to encompass cross-cultural assessments. Parallel forms and more flexible variations of them available under CTT assume that

means and standard errors are the same across language versions. It is not surprising that these would vary across countries because the comparisons are made across different populations. One of the main aims of cross-cultural tests is to ultimately compare countries and examine the differences between them. Later, I turn to LTT and examine statistical techniques used to establish equivalence amongst various tests under that theory. I use the literature to draw a distinction between ‘linking’ and ‘equating’ and argue on conceptual grounds that two language versions of the same test cannot be equated; they can only be linked. Language versions of international assessments are however not linked nor equated. They are translated and adapted following a very rigorous process. The last section of this chapter describes the international guidelines and the various methods used for translating and adapting tests. Finally, issues around these procedures and possible sources of bias associated with the methods are discussed.

3.1 Equivalent forms of an assessment under CTT

Developing equivalent forms of the same assessment was initially associated with the need to test the achievement of various groups of the same population with the same degree of reliability when administration of the same test at the same time was not possible for various theoretical and practical reasons. Theoretical reasons included eliminating the effect of memory in repeated measures or comparing the achievement of the same sample in pre-test/post-test situations. Practical reasons included the impossibility of carrying out the assessment at the same time or ensuring that the same test is equally reliable over years. These tests were often referred to as ‘parallel tests.’ The assumption was that if two tests are parallel then scores will be interchangeable across versions and comparisons can be made directly. The tight association between parallelism or equivalence and reliability has shaped the various definitions of parallel tests into very statistical definitions.

3.1.1 Parallel forms

One of the earliest descriptions of parallel forms was suggested by Gulliksen (1950) and was framed within CTT. In a rather casual way, he stated that “two tests are ‘parallel’ when ‘it makes no difference which test you use’” (p.11). In other words, individual i would have the same true score irrespective of the test form administered and the standard deviation of errors would be equal on both tests. This definition is a rather theoretical one. Gulliksen suggested a more practical definition, which is useful when dealing with empirical data: two tests are parallel when the means of their observed scores, the observed standard deviations, and all inter-correlations are equal.

Horst (1966) suggested a stricter definition under CTT. Parallel forms are tests which result from the selection of representative samples of unidimensional items from a universe of items covering a particular domain of measurement. Horst emphasized the necessity of having a one-to-one correspondence selection rather than a random selection technique; the latter not capable of guaranteeing that the total population of items selected is comprehensive enough of the construct to be measured. Horst suggested some additional criteria to define parallelism by pure logical reasoning, the most relevant of which are summarised below.

Two parallel tests are characterised by:

- an equal number of items in both tests
- equal sums of item difficulties given a particular sample of people
- equal means of observed scores between the two tests
- same power of discrimination of the tests
- equal means of item variances
- equal item covariances

Two years later, Lord and Novick (1968) formalised the definition of parallelism using mathematical formulae under CTT. They defined ‘parallel measurements [as ones which] measure exactly the same thing in the same scale and, in a sense, measure it equally well for all persons’ (p.48). And translating this definition into a more algebraic form, Lord and Novick wrote: ‘Distinct measurements X_{ga} and $X_{g'a}$ are called parallel measurements if, for every subject $a \in \mathcal{P}$, $\tau_{ga} = \tau_{g'a}$, and $\sigma(E_{ga}) = \sigma(E_{g'a})$ ’ (p.48) (X being the observed score, g and g' being given tests (or items), a being an individual in population $\mathcal{P}$ and τ being the true score (expected score), $\sigma(E_{ga})$ is the standard deviation of error-score E of individual a on test (item) g).

In simpler language, parallel tests are ones where, for individual a sitting both versions, the observed scores on both tests are equal, the true scores on both test forms are equal and the standard deviations of the error-scores are the same. Since true scores are not directly measurable, the authors derived a theorem based on previous definition: ‘Distinct measurements X_{ga} and $X_{g'a}$ are parallel measurements in $\mathcal{P}$ if and only if in every subpopulation of $\mathcal{P}$, $\mathbb{E}(X_{g*}) = \mathbb{E}(X_{g'*})$ and $\sigma^2(X_{g*}) = \sigma^2(X_{g'*})$ (p.48).’ That is, the expected values are equal, and the observed-score variances of parallel measurements are equal. It is

worth noting that Lord and Novick's fundamental theory was about a set of items in a test and the discussion about a set of tests is by analogy.

The above definitions framed within CTT seem straightforward and convenient; parallelism holds if the above conditions are satisfied. However, none of these definitions can be extended to encompass the case of cross-cultural assessments. Gulliksen's (1950) definition (i.e. equal observed scores, standard deviations and inter-correlations) cannot be extended to an international context where populations are different and hence observed scores and standard deviations are expected to vary in the different countries. The question in this context is rather whether differences in scores are due to variations in abilities only or whether there is (an)other factor(s) causing the differences.

Similarly, Horst's (1966) criteria cannot all be satisfied in international assessments for the same reasons. Apart from the first criterion which sets out that the number of items in parallel versions of a test should be equal, the rest of the criteria (such as the mean difficulty, mean scores, variances, etc.) are expected to vary across countries. Lord and Novick's (1968) definition is also inadequate as it involves equal expected values (mean of observed scores) and equal observed score variances. These are expected to differ across countries.

Parallel forms are the most 'equivalent' forms under CTT, but as we see, we can reject the notion that language versions of international tests are parallel forms. The next section presents variations of the definitions above within CTT which relax some of the criteria. Could any of these 'equivalent' forms describe the case of cross-cultural tests?

3.1.2 Levels of equivalence

Graham (2006) suggests a hierarchy of models of parallel forms based on the work of Lord and Novick (1968) and Raykov (1997a; 1997b). He constructs a continuum of decreasing parallelism starting with the ‘parallel model’, followed by the ‘tau-equivalent model’ and the ‘essentially tau-equivalent model’. Graham ends the continuum with ‘the congeneric model’ – a model not described in Lord and Novick’s (1968) book. According to this hierarchy, the parallel model is ideal, yet it is the least flexible where all items measure the same latent variable, on the same scale, with the same degree of precision, and with the same amount of error (Raykov, 1997a, 1997b). True scores of all items are assumed to be equal to one another and all error scores are equal across items. Under CTT, this can be written as: $X_{ik} = T_i + E_i$ (k = item; i = individual). The tau-equivalent model is slightly less restrictive than the parallel model; it allows for individual error item variance: $X_{ik} = T_i + E_{ik}$. The essentially tau-equivalent model is even less restrictive and is the only one that explicitly introduces the criterion of a common latent variable. In the essentially tau-equivalent model, all items measure the same latent variable and on the same scale. Not only might the amount of error vary from one item to another but also the degree of precision may vary from one item to another: $X_{ik} = (\alpha_k + T_i) + E_{ik}$.

The congeneric model is the most flexible of all four models. In congeneric tests, all items measure the same latent variable; however, measurement could be made on different scales, with potentially different degrees of precision and possibly different amounts of error: $X_{ik} = [\alpha_k + \beta_k (T_i)] + E_{ik}$ (where α_k is an additive constant introducing flexibility in precision, β_k a multiplicative constant introducing flexibility in the number of scales and E_{ik} showing differences in error per item).

The previous definitions were mostly focused on the statistical properties of the tests without much attention to what is measured. The latent variable, that is, the construct, is represented only as one of the criteria in the two ‘less restrictive’ models. However, it is important to note that, in the context of international assessments, restricting the comparison to the statistical level may lead to flawed interpretations. Two tests could hypothetically have similar statistical properties such as same observed mean score and same standard error yet they could be measuring different constructs. Would these tests be considered equivalent? If yes, what is actually ‘equivalent’ across these nations? Is it a case where Population A (of Nation A) is as proficient in reading as Population B (of Nation B) is proficient in algebra? Is that comparison meaningful? The importance of having the same construct measured is elaborated further elsewhere in the chapter (Section 3.2.2).

The hierarchy proposed by Graham (2006) may be useful in the way different language versions of the same test administered to different populations with fundamental cultural differences such as in the case of PISA exams can be conceptualized. Since the strictly parallel model does not hold in the case of cross-cultural comparisons for the reasons laid out earlier, then allowing a level of flexibility might allow a better conceptualisation of these examinations. It would be difficult to claim that comparisons of the kind made in international assessments could be carried out with the same amount of error or a single level of precision. Thus, looking at PISA tests as essentially tau-equivalent or even congeneric tests seems more reasonable especially that in both models, the importance of having a common measured trait is articulated.

However, both models apply to tests administered within the same population and their definitions do not extend to cross-cultural comparisons. This aspect is particularly

important under CTT because comparisons are sensitive to sample selection and population distributions. This means that none of the models of equivalent tests available under CTT describes well the case of language versions of the same test. In the next section, I examine statistical techniques available within LTT that permit the direct comparison of scores across countries and are less sensitive to population distributions and sampling. In addition, I discuss the extent to which these techniques are applicable in the context of international tests.

3.2 Equivalent forms of an assessment under LTT

Linking is a statistical technique of mapping scores of two or more tests on the same measurement scale in order to allow direct comparisons across different tests. This is possible under LTT because IRT and conditioning procedures used are not affected by population distributions (Holland & Wainer, 1993). In this section, various levels of linking are presented and a distinction between linking and equating is discussed. Later, the possible classification of international assessments is examined. This section ends with an emphasis on the importance of having equivalence at the level of the construct.

3.2.1 Types or levels of linkage?

Linn (1993) classified types of linking in terms of stringency of requirements. According to him, only strong methods of linking can guarantee sound inferences that assume the interchangeability of scores. Weaker forms of linking can lead to other types of inferences, however, these are contingent upon the context, the time and the groups assessed. Holland and Dorans (2006) elaborated on Linn's (1993) classification and developed a framework consisting of three main categories of linking. The three categories are ordered along a

continuum of increasing rigor and based on the purpose fulfilled by each of these comparisons: (1) predicting, (2) scale aligning, and (3) equating. Predicting aims to estimate a candidate's score on one test from data gathered about this candidate. Scale aligning links the scores of two different tests through a common scale. Equating produces two tests with interchangeable scores.

Newton (2010) disagreed with this hierarchy and criticised the deficit rhetoric implied in the continuum suggested by Holland and Dorans (2006) whereby the evaluation of the quality of a particular link is based on departures from the 'equating' golden standard. According to him, equating and linking differ in type rather than in degree. Equating is a particular instance of comparing two examinations designed to the same framework and syllabus and, measuring identical constructs (Newton 2005; 2007). In other words, equating occurs when comparing two versions of the same test covering the same underlying construct where the only difference lies in the specific questions asked (Newton, 2007). Linking, however, refers to comparing examinations developed for different syllabi, disciplines, cultures, etc.

Often, versions of examinations measuring different constructs, at different times and covering different syllabuses, are linked (e.g. GCSE Chemistry examinations from two different examination boards). According to von Davier (2010), a link between two tests is only suitable when it does not impact test-takers negatively. She added that the appropriateness of establishing a link between two tests relates back to the extent to which each of these examinations conforms to the conditions imposed on educational assessment and is thus contingent upon the following five factors: (1) the same construct is measured in both tests; (2) same reliability on both tests; (3) the link is fair for both tests; (4) linking does not jeopardise the alignment of the test construct, testing instrument, goal and use for

both tests; (5) the predictive validity of either of the two tests remains constant. Von Davier (2010) adopted the ‘deficit rhetoric’ criticised by Newton (2010) and argued that equating fulfilled all these conditions together while other forms of comparability satisfied each of them to different degrees.

3.2.2 Linking language versions of the same assessment

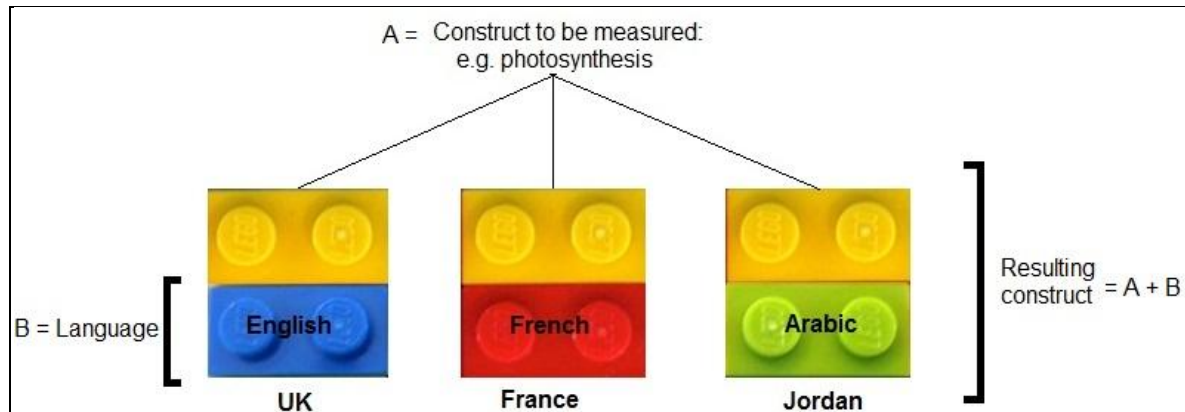
In keeping with Newton’s (2010) distinction between linking and equating, that is, adopting the view that the distinction between linking and equating lies in the nature of the linkage made and not in its degree, it is worth wondering whether language versions of the same international test are linked or equated versions of the same test. The answer is simple. Based on Newton’s (2005; 2007) criteria, given that these tests are administered to different groups of people then this cannot be a case of equating. Hence, caution should be exercised when interpreting results.

However, for language versions to be considered linked versions there should be common people and common items between the versions. This is particularly problematic in the case of international assessments where tests are translated. A technical challenge arises because there are no common people between any two countries taking the exams. Besides, the items are translated versions but not exactly the same. To best illustrate this point, consider Figure 3.1 below.

In Figure 3.1, Lego parts are used to illustrate the different parts of a science construct. In this figure, the same hypothetical test covering a given construct in three languages is compared across three countries (UK, France and Jordan) administering the test in three languages (English, French and Arabic, respectively). For the sake of simplicity, let’s assume that the construct of interest is composed of two parts: Part A, which is the

intended construct (in this case photosynthesis) and Part B, which is the language part. When comparing the three tests, let us suppose that Part A is constant while Part B varies by definition. In international assessments, the constructs are similar to the ones represented in Figure 3.1; they have a part in common yet they are not exactly the same.

Figure 3.1: Lego representation of the same construct in three languages



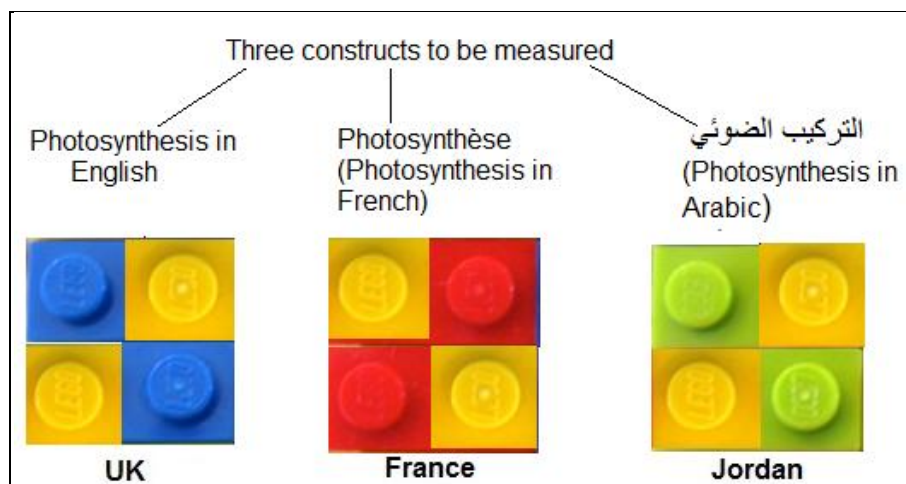
Technically speaking, virtually any two examinations could be linked as long as the tests have common people and common items (Newton, 2005). However, Newton (2005; 2010) emphasised the importance of making appropriate and meaningful linkages. Comparability is justifiable ‘to the extent that linked scores are more adequate and appropriate for achieving a particular purpose than non-linked scores’ (Newton, 2010, p.55). According to Newton (2005; 2010), two assessments can always be linked meaningfully if they have a construct in common. Inferences about students at linked scores should be limited to this common construct. Newton gave the example of linking psychology and sociology tests in terms of a common construct: skills in social sciences. In the context of international assessments, it is worth investigating whether linking scores across language versions is meaningful. In chapter 2, I argued that language is part of the construct, be it intended or unintended by the examiner. In this section, I argue that when attempting to link scores across countries in international assessments, not only is there no common people. There

are no really common items between any two countries administering different versions of the test either.

This can be portrayed in Figure 3.2 by amending the Lego representation of Figure 3.1. In Figure 3.2, the construct measured (in this case, photosynthesis) is influenced and shaped by the language used to express it and hence gives rise to a new construct in each country. Accordingly and based on Newton (2005; 2010), the comparison of scores across language versions of the same test is not meaningful. I would nuance this claim by stating that the comparison would only be meaningful to the extent that the various countries are worth comparing.

For instance, it may be reasonable to compare how 15-year old students achieve in science in two Western developed countries such as UK and France and less justifiable to make such a comparison between 15-year old students of two countries with distant cultures and different economic wealth (e.g. UK and Jordan). It is important to note, that some research questions aim to investigate differences across cultures and insightful findings can result from such studies. However, it is questionable whether using the same test in different languages is the best methodology to answer such research questions.

Figure 3.2: Lego representation of three constructs of the same underlying concept



Grisay et al. (2007) raised similar concerns about whether PISA is a suitable instrument to compare countries in the Middle East and part of Asia and suggested that these countries may need to devise a more appropriate and relevant test. According to Poortinga (1995), translation is not always capable of transferring the conceptual differences intrinsic to a particular language into another one. Bonnet (2002) comments on the comparability of item difficulty across language versions of the same test by stating that it is hard to imagine that when taking cultural and language differences of candidates into account, a test translated into Korean would have the same level of difficulty than ‘the same test’ generated in Finnish. Discussing translation from English to Welsh, Jones (1998, p.179) claimed that: ‘no amount of painstaking translation will quite eradicate the Englishness of thought which underlies it.’ This supports the argument that language is an intrinsic part of the construct being tested and is hard to separate.

In this section, I discussed the distinction between equated and linked versions of the same test under LTT. I argued that language versions of the same test can at best be considered linked versions of the same test. I also claimed that the meaningfulness of the linkage established and the significance of the interpretations ensuing from any comparisons across countries are determined by the nature of the research questions and the degree of similarity between the various countries. The above discussion was carried out on conceptual grounds. However, international assessments are only linked longitudinally within each country; i.e. across cycles (OECD, 2009b), not across language versions. In PISA, rigorous scrutiny of the equivalence of items and tests between the source version and the national version is achieved by expert judges and trained translators (see Chapter 5, Section 1e) but no subsequent equating or linking of language versions of the same test is performed. Similar practices are carried out in translating and adapting intelligent tests such as WISC-III (see Georgas et al, 2003).

The fact that equivalence in international tests is maintained using expert judgement does not imply that linking international tests would be carried out without any methodological challenges. Sireci (1997) discusses a number of methodological pitfalls when linking language versions of the same test using IRT. He highlights the lack of construct invariance as being a major challenge because of the lack of common items in reality. However, the current practice in international assessments raises concerns over the extent to which expert judgement is reliable enough to ensure high levels of equivalence of language versions of tests. Empirical evidence of differential item functioning in PISA (Bonnet, 2002; Grisay & Monseur, 2007; Grisay et al., 2007; Huang et al., 2014; Kreiner & Christensen, 2014; Le, 2009; Oliveri & Ercikan, 2011; Xie & Wilson, 2008) suggests that some items were biased for or against particular countries. This bias could have originated from the way tests were translated and adapted. The following section articulates the meaning of test adaptation and explains the distinction between translation and adaptation. Later, three commonly used methods of test translation are presented and sources of bias of test adaption are discussed.

3.3 Test adaptation

With the increasingly global economy of the twenty-first century and the rise of international talent exchange, instances of comparing student achievement across countries has become ever more in demand. Exams are used to provide credentials at international level and interest in cross-cultural research has increased (Hambleton et al, 2005). This has led to an increasing need to translate exams into different languages and adapt them for different cultures. Translating a test from the source language to another language, the target language, involves the conversion of the content of that test from the source language to a target language. Assessments are well translated if both linguistic and

psychological characteristics of the source text are retained in the target version (Hambleton, 2005; van de Vijver & Poortinga, 2005). The linguistic characteristics include semantic meaning, readability, comprehensibility and style. The psychological features consist of ways to preclude bias which may occur when assessments are administered in different cultures (Hambleton, 2005). This additional level of translation moves the process of test translation from literal translation to test adaptation.

This section describes various options for test translation. Their application depends on the nature of the test and the extent to which it is translatable. A brief description of PISA test adaptation procedures is also provided. Later, sources of invalidity of test adaptation standards are discussed. These include cultural and language differences, technical issues related to designs and methods adopted in the process and factors affecting the interpretation of results.

3.3.1 Options for test translation

Depending on the extent to which an instrument is readily translatable, three options are available in test translation: application, adaptation or assembly (van de Vijver & Poortinga, 2005). Each of these options is described below.

3.3.1.a Application option

Some assessments are highly translatable into some other languages. This occurs when the two cultures are similar. Psychological features can be easily retained in the target version so bias is less likely to be present. The application option (van de Vijver & Leung, 1997) consists of a literal translation of a test. This option is simple and cost-effective and many cross-cultural studies use it despite it not always being the optimal option.

For instance, the item below is an arithmetic question taken from TIMSS 1995 (Maths test, Population 2). Arithmetic questions consist mainly of number operations which symbols are universal and have less words. This is why they lend themselves well to the application option; that is, literal translation.

*Figure 3.3a: TIMSS 1995 Maths item (Population 2)*⁷

M8. Multiply: $0.203 \times 0.56 =$

Answer: _____

Reproduced from TIMSS Population 2 Item Pool. Copyright © 1994 by IEA, The Hague

This item, targeting lower secondary students (Years 7 and 8) can be translated literally into French (see below). Assuming that curricula of the target French-speaking country covers the concept of multiplying decimal numbers by Year 7 or 8, it is not likely to see any bias associated with this item.

Figure 3.3b: Application of TIMSS 1995 Maths item (Population 2)

M8. Multiplier: $0.203 \times 0.56 =$

Réponse : _____

3.3.1.b Adaptation option

Some assessments are moderately translatable; literal translation of some parts of the assessment into the target language are possible without introducing bias while other parts

⁷ See the following website for released TIMSS Grade 8 items and scoring
<http://www.education.gov.za/LinkClick.aspx?fileticket=gWrcovqUMI%3D&tabid=666&mid=1861>.

may have to be altered to ensure that linguistic and psychological criteria are equivalent in both versions. Test adaptation requires ensuring that concepts, words and expressions used in the adapted version are culturally, psychologically and linguistically equivalent to those in the source version. A typical example is when a currency is involved. For instance, the item below is a mathematics item (TIMSS 1995, Population 2) involving American dollars (\$).

Figure 3.4a: TIMSS 1995 Maths item (Population 2)⁸

R13. Mr. Lewis had \$360. He spent $\frac{7}{9}$ of it. How much money did he have left?

Answer: _____

Reproduced from TIMSS Population 2 Item Pool. Copyright © 1994 by IEA, The Hague

If used in the UK, this item would need minimal adaptation: the conversion of American dollars to British pounds (£). If used in France, more adaptation will be needed. In addition to the currency change to Euros (€), the item has to be translated (See Figure 3.4b).

Figure 3.4b: Adaptation of TIMSS 1995 Maths item (Population 2)

R13. Mr. Louis a €360. Il dépense $\frac{7}{9}$ de son argent. Combien d'argent lui reste-t-il?

Réponse : _____

⁸ See the following website for released TIMSS Grade 8 items and scoring
<http://www.education.gov.za/LinkClick.aspx?fileticket=gWrcoxvqUMI%3D&tabid=666&mid=1861>.

Note that the choice of the present tense in the translation into French was deliberate. The past tense that would be grammatically correct in this case would include the use of *passé simple* in *dépensa* and *resta* (*Mr. Louis avait €360. Il dépensa $\frac{7}{9}$ de son argent. Combien d'argent lui resta-t-il?*). The use of such tense may increase the reading demands of the item in French as, unlike what the name implies, this past tense is not the simplest in French in terms of verb conjugation.

3.3.1.c Assembly option

Some assessments are poorly translatable to other languages; normally because the source and target populations differ on many aspects, such as the way they define a particular construct. For instance, the concept of intelligence varies across cultures. Western psychologists associate intelligence with the volume of knowledge that someone has and the speed with which s/he can generate answers to questions. However, in certain communities in Uganda, intelligence is perceived as wisdom, slow contemplation of matters and giving the correct response (Wober, 1974). In Kenya, Kokwet mothers define intelligent children as children who value family ties and are well-behaved (Mundy-Castle, 1974 quoted in Segall et al., 1990). In this case, literal translation or mere adaptation is not sufficient to ensure equivalence between source and translated versions. A new assessment is assembled to preclude potential bias and guarantee equivalent coverage of the construct. It is clear that this option makes direct comparison of scores very challenging. But these assessments could be useful in showing the disparity between cultures in relation to the construct measured.

PISA adopts a rigorous design of double translation to limit cultural bias (see Chapter 5, Section 1e). Choosing the ‘adaptation option’ over the ‘application option’ or ‘assembly option’ presupposes that the cultures in questions are distant enough not to choose the first

option, i.e. the mere translation of the tests, but close enough not to have to resort to the third option and assemble an entirely new assessment covering an equivalent construct. This decision might be adequate when comparing two developed European countries, such as France and Germany. However, it may not be adequate when comparing two culturally distant countries such as Germany and Japan, let alone comparing a developed with a non-developed country (for instance, comparing Jordan and UK). The degree of similarity between the countries compared determines to a great extent the nature and strength of equivalence as well as the possible types of bias that may emerge. The following section examines the various types of equivalence established in the context of test adaptation and discusses later types of potential bias.

3.4 Types of equivalence and of bias in adapted tests

In the previous section, I discussed the various options for adapting tests. In this section, I examine the various types of equivalence by presenting the taxonomy suggested by van de Vijver and Tanzer (2004). Later, I discuss the various types of bias that may threaten equivalence.

3.4.1 Types of equivalence

Lack of bias, irrespective of its type, is a pre-requisite for comparing scores across cultures and equivalence of adapted tests (van de Vijver & Tanzer, 2004). Strong claims about cultural differences in abilities are often made with equivalence of tests being assumed rather than demonstrated. Equivalence across cultures can occur at various levels. In decreasing order of equivalence, these are: scalar or full score equivalence, measurement

unit equivalence, structural or functional equivalence, construct inequivalence. The various types of equivalence are described below.

Scalar or full score equivalence is when instruments have the same measurement unit and the same origin across cultural groups. It is supposed to be the level with the greatest opportunity of cross-cultural comparisons. This level of equivalence is challenging to achieve especially when cultures are distant and languages are not related at all. In PISA, such type of equivalence may be established between the French versions of Belgium, Switzerland and France or the English version of UK, Canada, Australia and USA. Indeed, Grisay and Monseur's (2007) study suggests that countries with the same language tend to have only a smaller percentage of items with significant bias (2.4% amongst English versions and 2.6% amongst French versions).

Measurement unit equivalence refers to an instrument with the same measurement unit yet with different origins across cultural groups. The scales of each group should have the same metric. Van de Vijver and Poortinga (2005) gave the example of comparing different cultural groups using adapted versions of a test on extraversion and making gender comparisons within each group. If this test satisfied measurement unit equivalence then comparisons of gender differences could be made across groups but direct comparisons of individuals across groups would not be possible as the origin would not be the same on the different scales.

Structural or functional equivalence refers to instances when an instrument measures the same construct across the source and target culture(s). In the following I refer to this type of equivalence as construct equivalence. In a factor analysis, the source and adapted versions of a test should reveal the same factors at test level without necessarily having a one-to-one correspondence of variables at item level. Direct comparison of scores across

cultures would not be possible. Van de Vijver and Tanzer (2004) describe the opposite of construct equivalence as being instances of construct inequivalence where comparability would not be possible such as when ‘apples and oranges are compared’ (p.121).

A very rigorous adaptation process is employed in the development of PISA assessments to establish construct equivalence at the very least. That is, by limiting cultural bias, PISA claims to develop tests which measure the same constructs across all cultural groups. Moreover, complex psychometric techniques are used in the analysis of the data to ensure scalar or full score comparisons. The comparison of overall and subject-specific achievement across countries is carried out by comparing the national average achievement to the OECD average, set at 500 points with a standard deviation of 100 points (OECD, 2009b).

With the great number of participating countries involved in the studies, establishing the many types of equivalence is very challenging. Indeed, often construct equivalence is totally assumed (Poortinga, 1989). It is also worth noting that equivalence across cultural tests and bias in tests are judgements given about particular tests in specific instances of comparisons; bias and equivalence are not absolute properties of a given test (van de Vijver & Tanzer, 2004). A test could show bias when comparing France and Poland but show equivalence between France and UK. In PISA assessments, equivalence is established between the national version(s) and at least one of the two source versions. Establishing a more ‘absolute’ equivalence in PISA 2006 tests would not only entail reviewing the quality of the 87 national versions in 44 languages (OECD, 2009b) against both source versions. It would also mean comparing the 87 national versions amongst each other. This amounts to 3,741 comparisons to make in 44 languages!

Given that ‘absolute’ equivalence is nearly impossible to achieve, especially in large-scale international assessments, the presence of bias to a certain degree is inevitable. The section below discusses the various types of bias.

3.4.2 Types of bias

Various sources of bias exist; however, these tend to result in one or several of the four main types of bias: construct bias, method bias, item bias, and sample bias (van de Vijver & Poortinga, 1997; van de Vijver & Tanzer, 2004). Table 3.1 below is a reproduction of the updated version of this table (Hambleton et al., 2005, Table 2.2, p.43).

Table 3.1: Typical sources for three types of bias in cross-cultural assessment

Type of bias	Source of bias
Construct bias	- Dissimilarity in the definitions of the construct across cultures - Differential appropriateness of the behaviors <i>[sic]</i> associated with the construct (e.g., skills do not belong to the repertoire of one of the cultural groups)
Method bias	- Incomparability of samples (e.g., caused by differences in education, motivation) - Differential familiarity with stimulus material - Differential familiarity with response procedures - Differential response styles (e.g., social desirability, extremity scoring, acquiescence) - Differences in environmental administration conditions, physical (e.g. recording devices) or social (e.g., class size)
Item bias	- Poor item translation and/or ambiguous items - Item-related nuisance factors (e.g., item may invoke additional traits or abilities)

Construct bias refers to the partial overlap of a given construct between two or more cultural groups. Van de Vijver and Poortinga (2005) refer to studies carried out on filial

piety (e.g. Ho, 1996) and distinctions existing in the definition of this construct between the Chinese culture and the Western one. Filial piety in the Chinese understanding goes beyond the concept of obedience and respect towards one's parents (as defined in the west); it also includes provision of care and financial support to parents at an old age. This is a case of construct underrepresentation. This is likely to occur in PISA assessments where countries define curricular constructs differently. This point will be further elaborated later in Section 3.5.1.

Method bias includes two sub-types: instrument bias and administration bias. Instrument bias occurs when some features in the instrument are construct-irrelevant yet induce score differences. An example is the familiarity with the type of question. For instance, PISA uses items linked to a particular stimulus. This test format requires examinees to retrieve some of the answers from the stimulus. Students trained to answer questions within this type of format are better prepared to answer PISA questions than their counterparts who are less familiar with this type of questions.

Administration bias relates to miscommunications between examiner and examinee. This is more likely to happen in cross-cultural tests when the first language of examiner and examinee is different. Item bias, often referred to as differential item functioning in the literature, refers to anomalies in the items such as poor translation. Note that in Chapter 7, I highlight a nuance between the term item bias and differential item functioning. However, this distinction is not important at this point.

In this section, I have examined the various options of adaptation that can be adopted when developing cross-cultural tests. Test adaptation seems to be the option adopted by PISA despite the wide range of comparability existing between any pair of countries, from very similar cultures (e.g. Norway and Sweden; or UK and France) to very distinct ones (e.g.

Mexico and Japan; or UK and Jordan). I have also summarised the various types of equivalence established between any two adapted tests and have discussed the potential types of bias. Different types of biases threaten the validity of adapted tests to various degrees and emerge from different sources. In the next section, I discuss the sources of invalidity that can undermine test adaptation.

3.5 The challenges of the test adaptation

There are various challenges which may jeopardise the validity of a test. Hambleton et al. (2005) organised these sources of invalidity into three main categories: (a) cultural and language differences; (b) technical issues, designs and methods; (c) interpretation of results. The following section uses the same categorisation to discuss the sources of invalidity of adapted tests.

3.5.1 Cultural and language differences affecting scores

Scores are affected by variables sensitive to culture and language differences. These include: (1) the construct measured, (2) quality of communication in test administration, (3) familiarity with the test format, and (4) speededness.

3.5.1.a Construct equivalence

One of the most important aspects of equivalence across language versions of the same test is the use of equivalent constructs. Construct equivalence includes comparability of the conceptual and functional nature of the construct measured and the operationalisation of the construct in the cultures of interest (Harkness, 1998). For instance, the construct of ‘biological evolution’ might mean and be operationalised differently in various parts of the world. While scientists claim that it is the overarching framework for understanding

biology (Dobzhansky, 1973), in the USA and the Arab world learning human evolution is a contentious subject in science education (Vlaardingerbroek & El-Masri, 2006) as it is perceived to challenge religious beliefs. Hence, in many parts of the world, only selected areas (if any) of the theory are covered while in other parts of the world, the theory is being taught as the key paradigm for understanding all biology.

This differential coverage of a particular construct has wide implications on international assessment. The concept of biological evolution appears as a unit amongst the PISA released units (Unit S472⁹). However, this unit was not administered in PISA 2006 cycle. Nevertheless, the concept of evolution can be traced in at least one administered unit in PISA 2006 cycle – the Grand Canyon (S426) – where one of the questions invoked fossils. Studying fossil records is at the heart of biology units discussing the phenomenon of extinction. This phenomenon is tightly related to concepts like natural selection and speciation. Students within educational systems that emphasize biological evolution will be at an advantage when solving such items compared with students who are less familiar with the concept and may have been taught creationism.

3.5.1.b Quality of communication in test administration

The quality of communication between examiners and examinees has a great impact on the validity of test scores. Minimizing ambiguity in test instructions is crucial and makes the process of adaptation simpler. Van de Vijver & Poortinga (1991) recommended using clear and self-explanatory instructions with little reliance on verbal communication. However, in assessments like PISA, the extent of clarity of instructions is likely to vary with languages and the quality of adaptation across countries. One way of alleviating

⁹ Visit the following website for PISA released items and scoring: <http://www.oecd.org/pisa/38709385.pdf>.

communication problems is to avoid elements of surprise in the test format or type of items (Baird et al., 2013). Familiarity with test format is discussed next.

3.5.1.c Familiarity with test format

Differential familiarity with test format may become a source of invalidity in cross-cultural comparisons of test results. Different cultures are more or less familiar with particular item formats. For instance, educational systems historically based on the British system are more familiar with essay type questions while educational systems based on the American model are more acquainted with multiple choice type of questions. Ways to avoid invalidity of test results include adopting a single item format that is common for all cultures sitting the test or diversifying and balancing item formats within a test (Hambleton, 2005).

Using one type of item format might result in narrowing the intended construct to the aspects that can be measured using the common item format. International assessments have resorted to including a variety of types of items. PISA tests comprise multiple choice items, short closed-ended responses, open-ended responses (See Chapter 5, Section 1a). Nevertheless, many countries are introducing PISA-like items in their curriculum (e.g. reading test in Norway and science tests in Japan). This might threaten the validity of PISA because students might develop ‘test wiseness’, in other words, they might become skilful at solving this type of items because of increased item familiarity. Studies carried out on coaching in the context of aptitude tests (Allalouf & Ben-Shakhar, 1998; Bunting & Mooney, 2001; Griffin, Carless & Wilson, 2013; Kulik, Bangert-Drowns & Kulik, 1984, Powers & Rock, 1999) have not been conclusive in this regard. Several studies pointed to a relative increase of scores following extended periods of coaching (Bunting & Mooney, 2001) especially for higher ability individuals (Griffin et al., 2013).

3.5.1.d Speededness

Taking a test quickly is an acquired skill more or less emphasized in different cultures. For instance, evidence points to ethnic bias due to test speededness on the Scholastic Aptitude Test (SAT) in USA. Differential speededness was observed on the SAT verbal test where African American and Hispanic examinees were disadvantaged compared to their White counterparts of comparable ability (Dorans et al., 1992; Lawrence, 1993). These tests are often biased against poor readers and unless answering a question quickly is part of the construct measured, it is necessary to limit test speededness by allowing enough time for the average examinee to complete all items. Normally items appearing at the end of a test show more bias.

In PISA, one of the aims of field testing items prior to their administration is to make sure that enough time is allocated for completing the items (See Chapter 5, Section 1). However, Eivers (2010) compared three versions of PISA tests with the English source version. He found that the Finish versions were on average 8% longer than the English versions while the Irish versions were 11% longer and the German versions were 17% longer. Longer tests require more time to read; however, all countries are given exactly the same time (two hours) to cover the test. This suggests that the German versions are more prone to be affected by speededness than do the Irish, or the Finnish or the English versions.

The above factors may vary across cultures and hence manifest themselves distinctively in different language versions of a single test. The same aspect may turn an item easier in culture A yet harder in culture B. When combined together, they make the predictability of equivalence across various versions of the same test administered in different cultures very complex. The systematic difference between source versions and adapted version resulting

from the poor level of equivalence between the two versions is likely to affect scores differently introducing systematic bias and making the possibility of direct comparisons very weak. Cultural differences affecting scores are, however, not the only factors invalidating the process of test adaptation. Issues associated with the techniques, methods and designs used in the process can also threaten the validity of adapted tests. The following section discusses these factors.

3.5.2 Technical issues, designs and methods

According to Hambleton (2005), there are at least five technical factors that can influence the validity of tests adapted for use in other languages and cultures: (a) the nature of the test, (b) the selection and training of translators, (c) the process of translation; (d) judgemental designs for adapting tests; (e) data collection design and data analysis for establishing test and item equivalence.

3.5.2.a *The nature of the test*

Assessments can be more or less translatable into different cultures. This depends on the nature of the test, the content, the test format, the relative cultural distance between source and target culture(s). Malda et al. (2010) created two test versions assessing short-term memory, attention, working memory, and reasoning for Afrikaans and Tswana children in South Africa. Results suggest that better performance was associated with content familiarity. For instance, Tswana children scored better on items referring to soccer while Afrikaans children performed better on similar items invoking rugby. International tests, such as PISA, are developed while taking into account these matters at the outset to make the adaptation process less problematic. The choice of various features, such as item

format, stimuli and language, in the source versions is made in a way to minimize cultural bias and make it easier to adapt (OECD, 2009b).

3.5.2.b Selection and training of translator

Translators should be carefully selected when adapting tests. It is not enough to find qualified translators who are equally proficient in both the source and target language. Translators should also have a background in the subject translated and have knowledge of test and scale construction. Otherwise, they could unintentionally introduce elements that make the test easier or more difficult in the adapted version. For example, they might unknowingly insert clues within distractors making them easier to rule out. Below, an example of a released PISA science item¹⁰ (S508Q03 – Genetically modified crops unit) illustrates this point.

Figure 3.5a presents the English version of an item administered as part of PISA science 2006 (Refer to Appendix to 3.1¹¹). Item S508Q03 was a simple multiple choice question with D being the key (that is, the correct answer) and A, B, C being distractors (that is, wrong options). The language used in the item was simple with limited use of scientific terms. The term genetically modified crops and the acronym GM were defined in the stimulus, that is, in the text preceding the item. The readability of the item did not seem to pose any potential problems to a 15-year old student.

¹⁰ See the following website for released PISA items and scoring: <http://www.oecd.org/pisa/38709385.pdf>

¹¹ See also Appendix to 3.1 for the English, French and Arabic versions of all PISA released items discussed in the thesis.

Figure 3.5a: PISA science 2006 multiple choice item

Question 3: GENETICALLY MODIFIED CROPS

Corn was planted in 200 fields across the country. Why did the scientists use more than one site?

A So that many farmers could try the new GM corn.
B To see how much GM corn they could grow.
C To cover as much land as possible with the GM crop.
☒ D To include various growth conditions for corn.

The adaptation of this item into French was not problematic. The language remained simple in the question as well as in the distractors and the key. That was the case with the Arabic version except for the key (i.e. option D). Figure 3.5b below presents the adaptation of the key from English into French and Arabic. The literal translation of the Arabic version of the key stands for: ‘*In the aim of testing the effect of different environmental conditions on the growth of corn*’. When comparing this to the English source version ‘To include various growth conditions for corn’, one can immediately notice the increase in complexity of the language in the Arabic version: a long nominal phrase which includes a high density of information. While all the Arabic versions of the distractors for this item were written in simple plain Arabic, the key was written in a rather scientific genre (See Chapter 2, Section 2.2.2.b). In this case, it is not the idiosyncrasies of the language that brings about the complexity; it is the poor quality of the adaptation. The key could be translated into a simpler form such as ‘لشمل ظروف نمو مختلفة للذرة’.

At first glance, one would think that by increasing the syntactic complexity of the language in the key and hence increasing the reading demands, the item would become more difficult in the country administering this Arabic version. In other words, this item would have a lower proportion of Arabic-speaking students getting it right. This however, remains a speculation that needs to be confirmed empirically. The item could well be made

easier because of the use of the scientific language in the key only and not the distractors and because of the introduction of the terms ‘aim’, ‘testing’, ‘effect’, ‘environment conditions’ which give clues to an experimental setting. Arabic-speaking students might be attracted by the smart-looking option. Unfortunately, this version is not the one used by Jordanians so it will not be possible to link the performance of Jordanians on this item to this version in Chapter 7. However this example illustrates well how untrained translators could unintentionally make the item more complex or introduce clues in the item and make it easier.

Figure 3.5b: Adaptation of item S508Q03 key into French and Arabic

English:	To include various growth conditions for corn.
French:	<i>Pour inclure diverses conditions de culture du maïs.</i>
Arabic:	بهدف فحص تأثير الظروف البيئية المختلفة على نمو الذرة.

3.5.2.c The process of translation

The *Oxford English Dictionary* online (2013) defines dialects as forms of a language used within specific regions or social groups. The problem of dialects within a language can become a threat to the validity of adapted tests. British English and American English use different vocabulary at times, such as pavement and sidewalk, lift and elevator, trousers and pants or aubergine and eggplant. For instance, Schittekatte et al. (2003) and Tewes (2003) highlighted the challenges faced in adapting the WISC-III intelligence test into Dutch and German respectively. Regional differences in dialects in those two languages posed major problems when constructing an unbiased version of the verbal items in the Netherlands and Belgium or Germany, Austria, and Switzerland.

Ideally, one could think of producing a version that would apply across dialects of a language. That could be done by ‘decentering’ the assessment; that is, by making it less

culture-based. PISA uses decentering as a technique: two different language versions act as source versions and are developed simultaneously (see Chapter 5, Section 1e). Also, countries are encouraged to contribute units. In PISA science 2006, countries were offered the choice to provide items following specific guidelines and in their national language after consultation with the PISA consortium. A total of 155 units were suggested from 21 countries. Most of these units were submitted in English, however, some items were developed in Dutch, French, German, Spanish and Swedish. Screening of these items in terms of their suitability for PISA 2006 reduced the number of items to only a handful: 13 units (19 items) of mathematics and 11 units (30 items) of science. Reading items used in PISA 2006 were all selected from PISA 2003.

3.5.2.d Judgmental designs for adapting tests

There are two common designs adopted in test adaptation: *forward translation* and *backward or back-translation*. The forward translation of a test consists of directly converting a test from the source language to the target language. Judgment over the quality of adaptation is carried out on the source and the target versions. This is considered to be the main advantage of this design. However, this design suffers from a high level of inference required from translators to establish equivalence of the two versions. Besides, judgement of equivalence is often made by bilinguals who may not be exactly equally proficient in both languages or may use their knowledge of both languages to judge equivalence (Hambleton, 2005).

A more frequently used design for adapting tests is back-translation. This design is considered to be more rigorous (Hambleton, 2005). It consists of translating the test from the source language to the target language and then back from the target language to the source language. Studies suggest that this method has identified various problems in the

adaptation process especially by minimizing cultural bias (Hambleton, 2005). However, this design has two main pitfalls. First, it does not allow a direct comparison between source and target versions of the test. Second, only versions in the source language are compared and hence any inappropriate aspects of the source language such as grammar and syntax could be retained.

PISA uses a double translation design where translation from two source versions, English and French, into the target language is conducted. Then the two translated versions thereabout are merged into a single adapted version that PISA test developers consider to have a higher quality of adaptation (OECD, 2009b).

Despite the rigour of the translation process, be it following a forward translation or backward translation design, evidence suggests that reviewers are incapable of flagging all problems in questions (Graesser et al., 2006; Pollitt et al., 2007) and this is why field testing of items remains essential in establishing solid grounds for equivalence across language versions. There are various ways to field test items in different languages. In the following, I present a review of the literature about various designs used in analysing data produced after field testing items. The case of developing tests in two languages is considered for simplicity. The endeavour becomes more challenging and complicated if equivalence is to be established across three or more languages such as in the case of large-scale international tests. When comparing two language versions of the same assessment, three designs are commonly used involving bilinguals or monolinguals. Each of these designs has its advantages and shortcomings. The following describes these three designs and discusses their respective strengths and weaknesses.

3.5.2.e Techniques for assessing equivalence at item and at test level

Equivalence of adapted tests can be analysed at item level or at test level using various data collection designs and methods. Establishing equivalence at item level is carried out using designs and techniques applied in item bias studies (Hambleton et al., 1991; Sireci & Allalouf, 2003). Various item statistics are compared across language versions of the test while controlling for any ability difference in the groups (see Hambleton & Kanjee, 1995). Items behaving differently are flagged and possible explanations for their differential functioning are proposed (see Grisay & Monseur, 2007; Huang et al., 2014; Le, 2009; Wu & Ercikan, 2006). Very often, poor translation and adaptation is associated with the emergence of bias (Grisay et al., 2007; Solano-Flores et al., 2009). However, the source of bias cannot only be attributed to flaws in the translation or the adaptation process. DIF techniques (see Chapter 7 for more details) can only point to inconsistencies in the behaviour of some items across groups. They are incapable of disentangling factors due to different educational systems or cultural differences from the quality of adaptation. Besides, given that DIF results from the interaction between some items in relation to the rest of the items, if all items are biased towards a particular group, no DIF would occur. Chapter 7 of this thesis analyses PISA data using DIF techniques and discusses the relevant literature more thoroughly.

DIF techniques identify poor equivalence at item level. Examining equivalence could also take place at test level. Three different data collection designs and methods are mostly used. The first one involves bilingual examinees, while the second two involve monolinguals. The three designs are described below.

- Using bilingual examinees

In this design, bilingual examinees take both the source and target versions of the test. The advantage of this design is that it makes use of a single sample. This means that the

examinee characteristics (such as age, gender, socioeconomic status, education background, etc.) are constant across language versions and hence scores are not affected by examinee differences (Hambleton & Bollwark, 1991). However, this design has two main weaknesses. First, this design assumes that bilingual examinees are equally proficient in both languages. Most bilingual examinees are significantly more proficient in one language in a given context (Sireci, 1997). Second, data of this design are not representative of any of the monolingual populations – neither the source nor the intended populations. Bilingual examinees tend to be, on average, substantially different on many levels from their monolingual counterparts (Hambleton, 1993). This limits the generalisability of the results and makes outcomes of studies based on this design useless in establishing equivalence across two specific monolingual cultures.

- Using monolingual examinees

The use of monolingual examinees has advantages in the sense that researchers could select a representative sample of the source language and the target language populations. Two designs are possible. These are described below.

The first design consists of administering the original form and the back-translated version to source language monolinguals. This design has two advantages. First, similarly to the previous design in which bilingual examinees are involved, using the same sample of participants allows researchers to control variables and eliminate the possibility of confounding differences in examinee characteristics with differences in scores (Hambleton & Bollwark, 1991). Second, the monolingual sample is representative of one of the populations of interest, the source population. However, this design suffers two main flaws. First, at no point is the target version analysed, despite the fact that the main purpose of the study is to evaluate the quality of the target version of the test. Second, if only a single sample of source language examinees is used, the results cannot be

considered independent because learning is very likely to occur between the administration of versions (the source and the back-translated versions). Although randomly assigning source-language examinees to one of the tests may help get around this problem, the first weakness of not analysing the target-version remains a major flaw in the design.

The second design of using monolinguals entails administering the original version of the test to a source language monolingual sample and administering the adapted version to a target language monolingual sample. This design has the advantage of using representative samples of both the source and target populations. In other words, findings based on this design are more prone to be generalisable to both the source and target populations. Also, unlike what is expected, and thanks to IRT models and conditioning procedures, no assumption about equal ability distributions has to be made (Holland & Wainer, 1993). However, this design makes use of samples of populations which differ to a greater or lesser extent. Conditioning procedures are helpful, yet when interpreting results, one should always keep the disparities separating the two cultures in mind.

The three designs for examining equivalent tests have their strengths and weaknesses and the selection of one design over the others is often made on pragmatic grounds such as related costs and availability of qualified people. Table 3.2 below compares the three designs and summarises some of the main strengths and weaknesses.

As shown in Table 3.2, the third design seems very appealing as the use of two monolingual samples would be representative of both populations of interest. However, in international assessments, populations often belong to very different cultures and the number of variables separating any two cultures is endless. Three main factors are particularly relevant and may affect the interpretation of results greatly. These are described in the next section.

Table 3.2: Designs for examining equivalence at test level

Design	No. of examinee samples	Language(s) of examinee sample(s)	Version administered	Population characteristics	Representativeness of data
Bilingual	1	A and B	source & adapted	Same	Neither
Monolingual	1	A	source & back-translated	Same	source population only
	2	A or B	source & adapted	Different	source & target population

3.6 Factors affecting interpretation of results

This section describes three main factors that normally affect the interpretation of results in cross-cultural settings: (1) differences in curricula; (2) student motivation; and (3) socioeconomic and political factors.

3.6.1 Differences in curricula

Different cultures follow different curricula that have more or less emphasis upon some subjects, content areas and skills. Interpretation of results should be conducted in light of these curricular differences. For instance, developing countries attribute great importance to health matters in the science curriculum (Baird et al., 2014). Assessments comparing science ability across those two countries should be built keeping that in mind.

Unlike TIMSS (see Martin et al., 2000), PISA claims not to be curriculum-based (OECD, 2009b), implying that curriculum scope would not advantage or disadvantage any participating country. However, this claim is highly questionable. Huang et al. (2014) found that curriculum cover is the main source of DIF when comparing Hong Kong and mainland China and between USA and mainland China.

Nardi (2008) described three possible relationships between PISA framework and the national curriculum: a commensurate relationship, an over-extensive relationship and an under-representative relationship. Baird et al. (2014) have nicely portrayed the three relationships in the Figure 3.6 below.

Nardi (2008) raised the issue of culture fairness in each case. According to her, in the commensurate case where the PISA framework and the national curriculum completely overlap, no culture fairness issue would arise. In the second case however, where the PISA framework is much broader than the national curriculum, students would be faced with questions about material they had not covered in school. For instance, in some countries, students work on graphs in mathematics yet are not trained to analyse and interpret them in a science context by the age of 15. The curriculum design might not be flawed and the emphasis might be simply on other aspects more relevant to that culture. Baird et al. (2014) mention also the case where some students have to study a larger number of subjects than their counterparts in other countries. This poses questions of fairness at the student level and leads to incorrect inferences about the effectiveness of the national educational system.

In the third case, where the national curriculum is broader than the PISA framework, Nardi (2008) finds that the issue of culture-fairness is associated with not giving the chance for students to demonstrate all what they know. Hence, irrespective of the relationship between PISA framework and the national curriculum, results of the former cannot be interpreted in isolation of the latter. Nardi (2008) suggests that only when the PISA framework is commensurate with the national curriculum does it not pose threat to the validity of the conclusions derived from the comparisons. According to her, it is only in this case, where international assessment can inform policy properly.

Figure 3.6: Possible relationships between PISA and national curricula



Source: Baird et al. (2014, p.67)

3.6.2 Student motivation

Not all examinees try their best to achieve a high score (van de Vijver & Poortinga, 1991). Student motivation varies in cultures and often relates to how high or low students perceive the stakes related to a test. Studies carried out on Swedish TIMSS 2003 indicate a positive and statistically significant relationship between motivation and performance (Eklöf, 2010). Hopfenbeck (2012) reported findings from interviews conducted with students in both Sweden and Norway after their participation in PISA (Eklöf et al., 2012). Students stated that they were doing their best despite the low stakes of the tests because they wanted to raise the profile of their country. This is not necessarily the case in other countries with a high testing culture such as in England (Hopfenbeck, 2012). Also different cultures set different achievement goals. For instance, Black South African aimed for the minimum required to pass a test (van de Vijver & Poortinga, 1991)

3.6.3 Socioeconomic and political factors

Needless to say, different cultures are embedded in distinct socioeconomic and political realities which confer different educational opportunities and different levels of access to resources (e.g. comparing results of a developed country with those of a developing

country). Although these are not the only factors explaining results, they do have a great contribution to the way a sample taken from a given culture performs on a test. For instance, environmental concepts appear in PISA science 2006 (Unit S114). Issues related to environmental protection might be covered by most national science curricula, yet students' daily practices in some countries, such as in developing countries, may not reflect these concepts. However, their counterparts in developed countries are embedded in a reality that reinforces their curriculum and offers them opportunities to engage with this concept on a daily basis (e.g. recycling garbage). Although environmental issues are not restricted to a particular culture, the level of awareness of those matters is however strongly linked to the extent to which a country is civilised. In some developing countries, concerns of having electrical power and political stability are more pressing. Developing countries are often coping with poverty, political unrest and health issues that environmental issues become neglected. Any comparison across cultures should take these variables into account.

The three factors discussed previously are the three most critical ones affecting the interpretation of cross-cultural comparisons. For instance, response rates in PISA have been criticised in several countries. While a lot of effort is put into selecting large samples of not less than 4,000 per country with very strict exclusion rules (OECD, 2009b), response rates varied across countries making some comparisons invalid. For instance, England's response rate was too low in PISA 2000 (Prais, 2003). OECD acknowledged later the lack of validity of results based on the data (OECD, 2010, p.1). The large number of factors challenging the process of test construction and adaptation has encouraged educational, psychological and testing agencies to develop standards and guidelines for test development and adaptations. The following section describes the most widely used ones.

3.7 Standards and guidelines for test adaptation

Two major documents defining sound practices and methods for developing, administering, using and adapting educational and psychological tests are available: (1) the *Standards for Educational and Psychological Testing* (1985; 1999) and, (2) the International Testing Commission (ITC) Guidelines for Test Adaptation (ITC, 2001). These are discussed below in relation to the level at which they are met in the PISA assessment process.

3.7.1 Standards for Educational and Psychological Testing

The *Standards for Educational and Psychological Testing* (1985; 1999) developed by joint forces of the American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) provide guidelines for sound development, administration and application of psychological and educational tests. These directions include two standards relevant to test adaptation of international assessments:

Standard 6.2: When a test user makes a substantial change in test format, mode of administration, instructions, language, or content, the user should revalidate the use of the test for the changed conditions or have a rationale supporting the claim that additional validation is not necessary or possible.

Standard 13.4: When a test is translated from one language or dialect to another, its reliability and validity for the uses intended in the linguistic groups to be tested should be established.

(Standards for Educational and Psychological Testing, 1985)

The standards above aim to ensure better quality of test adaptation and focus on maintaining the validity of adapted test versions. These standards are sometimes assumed

to be satisfied by PISA and are more or less clearly justified in the technical reports (OECD, 2009a; 2009b).

Standard 6.2 implies that revalidation of test use in the target cultures should be carried out unless solid justification is provided for their triviality in particular contexts. Construct validation in PISA was carried out on the context and attitudinal scales (see OECD, 2009a, Chapters 16 & 17) as well as on the data entry process (OECD, 2009a, Chapter 6). However, no validation was conducted on the subject-specific constructs. Extensive rounds of reviews and checks were carried out by experts in the relevant fields yet no field trialling of adapted reading, mathematics or science items took place.

Standard 13.4 was met by PISA to a large extent. National reliability indices for each subject and each domain were provided post PISA administration (OECD, 2009a, Tables 12.6-12.8). The validity was explored in terms of various aspects detailed in Chapter 14 of the technical report (OECD, 2009a), such as sampling, adaptation process, printing of materials, common requirements for test administration, quality monitoring, coding and data entry and processing. Detailed comments about departures of certain countries from the national data collection plan and the implementation standards were collated from various reports. Decisions regarding the quality and the validity of the data were described briefly in the same chapter (OECD, 2009a, Chapter 14). However, the validity described in this chapter related more to the development and implementation of the test as well the data analysis procedures rather than the validity of the subject-specific constructs across countries. The latter is totally assumed across cultures once the adaptation process has been proved sound and satisfying the PISA translation and adaptation guidelines.

3.7.2 The ITC Guidelines for Test Adaptation

The ITC produced a document with 22 guidelines for test adaptation – The ITC Guidelines for Test Adaptation (ITC, 2001). The document is organised into four parts: context, test development and adaptation, administration, documentation and score interpretation (see Appendix to 3.2). A lot of effort is invested in ensuring these guidelines are followed in PISA assessments. Careful and detailed instructions are offered at each stage of the test development, implementation and data analysis. Technical reports and data analysis manuals provide ample information. However, the difficulty of accessing the actual items as well as various gaps in the information released (Kreiner & Christensen, 2014) reduces the transparency of the process and keeps some questions about the validity of subject-specific constructs unanswered.

CONCLUSION

Various methodologies are available to establish equivalence across language versions of the same test. Psychometrics offer equating and linking techniques under LTT permitting the comparison of scores across tests to a degree allowed by the nature of the link. In this chapter, I contended that because language is part of the construct (as argued in Chapter 2) language versions of the same test cover different constructs. This makes linking scores across language versions of little value because the underlying comparison is based on different constructs; that is, the comparison does not rely on the same variable.

PISA equates national versions across administration cycles and not across language versions. Language versions are produced through double translation and adaptation of two source versions, the English and the French versions. This technique is believed to minimise cultural bias by decentering the assessment. The quality of the adaptation and

equivalence is carried out through various rounds of revisions by expert judges (see Chapter 5). For practical reasons, scrutiny of the adaption quality is limited to verifying the equivalence of each national version with one or ideally both source versions. For PISA 2006, establishing the equivalence across all language versions would have amounted to carrying out 3,741 comparisons in 44 languages. This is obviously extremely costly and time consuming.

Despite the rigour of the translation and adaptation techniques employed by PISA, empirical evidence points to the emergence of bias in different language versions of PISA. In this chapter, I discussed the various challenges entailed in translating and adapting tests. These problems stem from cultural and language differences as well as from the translation and adaptation techniques and designs employed. International associations such as AERA, NCME and ITC have developed standards and guidelines to ensure high quality and equivalence of adapted tests. However, often, the quality of equivalence in international assessments is judged subjectively by panels of experts. Studies show that agreement amongst judges is often very low (Graesser et al., 2006; Pollitt et al. 2007) which makes these judgements unreliable. Establishing cross-cultural equivalence of tests requires less subjective methods which can guarantee higher reliability. This however, does not resolve the issue with the validity of the comparisons across countries.

In Chapter 2, I argued that language is part of the science construct be it intended or not by the question setter. In this chapter I argued that language versions of the same construct are different constructs because the language element is different in each case. This means that comparing language versions of the same test suffers from a major conceptual flaw; that is, it lacks a common variable that forms the basis of the comparison across languages. This aspect is overlooked in international assessments. Countries are placed on the same scale

based on the overall performance of the selected national sample and based on their performances on the different subjects. Placing countries on the same scale is possible because PISA uses the Rasch model. In the next chapter, the conceptual underpinnings of the Rasch model are discussed.

CHAPTER 4

The Rasch Model

‘Ga toch fietsen!’
‘Go ride a bike!’

(Literal translation from Dutch)

INTRODUCTION

In Chapter 3, I used the literature on test translation and adaptation to stress the ample challenges that this endeavour entails. I outlined the various potential sources of bias and discussed how even with rigorous methodology of test translation and adaptation, bias is unavoidable at some level. I discussed the conceptual conundrum that statistical models and techniques for equating and linking tests add to an already complicated enterprise. An extended version of the Rasch model has been used to calibrate items and analyse data in PISA (Chapter 5) since its conception in 2000 until the last cycle administered in 2012¹².

There are strong arguments in support of using the Rasch model to analyse complex and large data sets. It is a simple yet an elegant model that places student abilities on a unidimensional scale of item difficulties. However, the model has several shortcomings, mainly when it comes to its rigidity. According to proponents of the Rasch model, the rigidity is a manifestation of the property of invariance that distinguishes this model from the rest of the IRT models such as the 2- and 3- parameter logistic (2-PL and 3-PL) models (Andrich, 1995). And hence, due to the property of invariance, this party considers that the model subscribes to fundamental measurement (Wright, 1997; Bond & Fox, 2001). At the other end of the

¹² The new contractor for PISA 2015, Educational Testing Service (ETS), adopted a 2-PL modelling approach to calibrate items of PISA 2015. See http://search.oecd.org/callsfortenders/Annex%20F1_FTAnalysisPlan-Cognitive_1.pdf.

continuum, 2-PL and 3-PL modellers consider that the Rasch model lies on unrealistic assumptions and fails to describe the data well (Traub, 1983). Andrich (2004) relates the debate to a paradigmatic conflict where both parties talk past each other. On the one hand, Rasch proponents contend that models should measure a particular variable in the data; and on the other hand, Rasch opponents argue that models should describe the data well.

This debate is of particular interest to my research as it often shapes interpretations of PISA results. In his book *Measuring the Mind*, Borsboom (2005) addressed the meaning of constructs within the field of psychology from three different stances: the formal, the empirical and the ontological. The formal stance relates to the model formulation and the interpretation of its formal terms. The empirical stance focuses on how the Rasch model handles data in actual research. The ontological stance discusses the extent to which psychometric concepts, such as person location on a scale, designate a real, existing entity or whether they are mere artefacts of research. Although Borsboom distinguishes these stances, he claims that in reality they work in coherence; researchers hold certain views that determine their choice of models and interpretations. I adopt Borsboom's approach to discuss my position in relation to the debate on measurement within the field of educational assessment while highlighting the complexity of comparing science abilities across languages when language is an inherent part of the construct.

This research compares three language versions (English, French and Arabic) of the same PISA science 2006 test in terms of item difficulty and demands. From the Rasch perspective, comparing language versions of the same test involves placing, on the same scale, the abilities of persons from two or more language groups on a given latent trait and the difficulties of items from two or more language versions of a test. The property of invariance in educational

assessment implies that, if the data fit the model, the comparison of two items is irrespective of the persons considered when making this comparison, and that the comparison of two persons is independent of the item used for the comparison (Rasch, 1961). In Chapter 6, I investigate empirically whether the property of invariance withstands the severe test of multiple language versions of the same instrument. In this chapter however, I examine the implications of this property in educational assessment and outline what can be learned from applying the Rasch model in international large-scale assessments.

This chapter is divided into four sections. In the first section, I discuss the formal stance of Rasch modelling by providing an overview of the model. In this section, I trace the Rasch model back to its origins; highlighting its relations to CTT and IRT. I then introduce the model in its dichotomous form and more general polytomous form. In the second section, I lay out arguments supporting the use of the Rasch model in this research by examining the model's empirical implications. I later explain the property of invariance and apply it to the PISA context. The implications of the property to educational assessment are discussed as well as its relevance to the overarching research question.

In the third section, I clearly state my position *vis-à-vis* the debate around scientific measurement in educational assessment and the role of the Rasch model in social science measurement. I raise questions about the extent to which psychological abilities are quantifiable. Moreover, I express concerns regarding comparing measurement in social sciences to that in physical sciences from an ontological perspective. In the fourth section, the different requirements of Rasch modelling are presented and major criticism of the model is discussed. The chapter ends by expressing my moderate stance with respect to the opposing

radical positions on Rasch modelling and by returning the discussion to the pragmatic epistemology described at the beginning of the thesis.

4.1 Overview of the Rasch model – the formal stance

For a long time, research in psychometrics and measurement theory has been based on CTT which focused on the observed total score of a person on a given test. The observed score is the sum of the true score measuring the person's location on a theoretical construct and an error score caused by the idiosyncratic circumstances peculiar to the particular test on a given occasion. In the fifties and early sixties, Frederic Lord and Georg Rasch's parallel work led to the emergence of a new theory– IRT, also referred to as latent test theory. IRT assumes that non-observable variables (also called latent constructs) can be measured using observable indicators (e.g. item/test scores). This relationship can be represented using a mathematical function (ogive function) relating a respondent's underlying ability (latent variable) and the probability of him/her giving a correct answer on a question in a test. Unlike CTT, IRT focuses on the interaction between the person and the item rather than the total score. This function can be characterised by three parameters: a difficulty parameter, a discrimination parameter and a guessing parameter.

The Rasch model, developed by Georg Rasch in the early sixties is considered to be the simplest of all IRT models because it consists of one parameter only to estimate the person location, that of the item difficulty. The Rasch model requires that all items are equally discriminating amongst the different groups taking the test and that random guessing is not involved. According to this model, accounting for the other two parameters, discrimination

and guessing parameters, contributes to the concealment of inconsistencies in the data. The following presents the formulae of the dichotomous and polytomous forms of the Rasch model.

The Rasch model for dichotomous items can be expressed using the following equation:

$$\Pr\{X_{ni} = x\} = \frac{e^{\beta_n - \delta_i}}{1 + e^{\beta_n - \delta_i}}, \quad (5.1)$$

In equation 5.1, X_{ni} is the response of person n on item i and x can take the values of 0 (incorrect) or 1 (correct), δ_i is the location of item i on a latent continuum, β_n is the location of person n on the same continuum.

For polytomous items, the Rasch Partial Credit model takes the form of:

$$\Pr\{X_{ni} = x\} = \frac{e^{\sum_{k=1}^x (\beta_n - \tau_{ki})}}{1 + \sum_{x=1}^m e^{\sum_{k=1}^x (\beta_n - \tau_{ki})}} \text{ (Andrich, 1978)}, \quad (5.2)$$

In equation 5.2, $X_{ni} = x \in \{0, 1, \dots, m_i\}$, is the maximum score for item i , and τ_{ki} is the k^{th} threshold location of item i on the latent continuum.

PISA uses an extended version of the PISA model which I will describe in Chapter 5 (Section 5.6.1). Given that PISA items comprise dichotomous and polytomous items, I employ both forms of the Rasch model to analyse the data in Chapter 6. In Chapter 3, I argued that no statistical model could fully address the research question because language is a confounding part of the construct of scientific literacy in adapted tests. However, within the IRT paradigm, useful information relating to the comparability of these tests can still be gained, even if no statistical approach can fully answer the research question.

4.2 Invariance in Rasch modelling – the empirical stance

International comparative large-scale educational assessments and measurement models emerged at the same time in the late 1950s and early 1960s and the development of one was often dependent on the other (Wendt et al., 2011). The publication of the seminal work of Georg Rasch, *Probabilistic Models for Some Intelligence and Attainment Tests* (Rasch, 1960), marked the move of these tests away from CTT to more probabilistic approaches which seemed to better justify and carry out measurement comparisons across different cultural groups. Today, the Rasch model is used in most international comparative large-scale educational assessments at least at some stage of item development or analysis (Wendt et al., 2011) and for deeper item analysis (Molenaar, 1995).

Until the 2012 cycle, PISA has applied Rasch modelling for data analysis, construct validity and reliability purposes (Wendt et al., 2001). TIMSS and PIRLS use Rasch modelling to analyse item-specific psychometric properties before IRT-scaling (Mullis & Martin, 2000). In this section, I first explain the property of invariance and illustrate its empirical implications in the context of this research. I then discuss the relationship of the property to fundamental measurement. Later, the Kuhnian perspective on the role of scientific measurement in science is described and its relevance to the function of Rasch modelling in social sciences is portrayed. The section delves into the debate on using the Rasch model as a measurement tool and states my position regarding this debate from an empirical stance.

4.2.1 The Rasch model and fundamental measurement

From a psychometric perspective, comparing people fairly on a given trait involves measuring this trait against a common scale. Measurement consists of constructing an instrument bearing a linear continuum divided into equal units from an arbitrary origin and allowing for a consistent mapping of an object's magnitude independent of the object considered. Hence, in keeping with the same psychometric perspective, comparing people's educational achievement entails the construction of a measurement tool suitable for assessing the trait of interest and independent of the people being involved; i.e. the instrument should be invariant across people. The requirement of invariance in social sciences was articulated by Thurstone (Engelhard, 1992) decades before Georg Rasch introduced his model (1961), where every item and every examinee can be located on the same scale. What distinguishes the Rasch model is that the criterion of invariance is algebraically built into it:

The comparison between two stimuli should be independent of which particular individuals were instrumental for the comparison. Symmetrically, a comparison between two individuals should be independent of which particular stimuli within the class considered were instrumental for the comparison (Rasch, 1961 p.332).

In the context of this research, invariance would imply that students with the same science ability should appear at roughly the same position on a measurement scale of science regardless of their home language. Translating Rasch's (1961, p.332) statement into the context of this research, invariance would imply that in each of the three countries of interest, the comparison between two items in a given language version, for instance English, is irrespective of the Anglophone person used to make this comparison and similarly, the comparison between two persons from the same language group, say English, is irrespective

of the item administered in English used to make this comparison. This is invariance manifested within the same country.

However, scores are compared across countries. This stretches the property of invariance to encompass comparisons amongst individuals across different countries and items across different language versions. It entails that the comparison between two language versions of the same item is irrespective of the person used for this comparison whatever language group s/he was selected from, and that the comparison between two persons from two different language groups matched on their ability is independent of the item used to make this comparison and irrespective of its language.

An immediate reaction would be – how can the ability of a French-speaking person be compared with that of an Arabic-speaking person using an item administered in Arabic or even English? This type of comparison is possible according to Rasch modellers, because the property of invariance is a property of the measurement tool and does not describe necessarily the data (Andrich, 1995). This point is picked up later in this chapter (Section 4.3) and clear positions related to measurement and invariance in social sciences are further discussed. It is, however, important to mention that the property of invariance along with many other properties made the Rasch model, according to its proponents, subscribe to criteria of fundamental measurement (Bond & Fox, 2001; Wright, 1997). This has various theoretical and practical implications. The rest of this section discusses those implications and starts with the theoretical one – that of drawing an analogy with the Kuhnian perspective on the role of measurement in science.

4.2.2 The Rasch model: Detecting anomaly in the data

The traditional goal of fundamental measurement is that of understanding a given variable or construct, estimating its location on a measurement instrument, quantifying it, assessing its behaviour and predicting it. In simpler terms, fundamental measurement is traditionally perceived as empirically understanding the behaviour of variables within the scope of given theories. Popper (1959) advocated a new scientific methodology based on the potential falsifiability criterion: empirical work should aim at disproving theories rather than confirming them. Popper's philosophy had an impact on how development in science was perceived at a time where inductivist and logical positivist approaches shaped the scientific method.

In 1962, Thomas Kuhn published his book *The Structure of Scientific Revolutions* where he described a new perspective on the development of scientific knowledge. In this book, he also suggested another purpose for measurement, one that he claimed to be the primary one; that of revealing anomaly in the data. Andrich (1988) refers to this perspective and contends that, given that the Rasch model subscribes to fundamental measurement, its main function should also be that of revealing inconsistencies in the data. In the following, I describe Kuhn's (1962) perspective on scientific revolution and the role of measurement in science and later discuss Andrich's (1988) claim. I then clearly state my stance within the debate.

Kuhn (1962) explored the philosophy and history of science and developed a new meaning for the word paradigm which describes a framework of understanding underpinning a science. Paradigms often set the theories used to understand the data, the methods employed to collect them and the plausible interpretations that could be reached. Paradigms define the research agenda, and usually, proponents take them for granted and often struggle with ways to make

data conform to them. Very often different paradigms coexist and proponents of each seem to be talking through each other because they are not detecting the fundamental disparities in the philosophy leading their thinking. They are often not even addressing the same problems.

Research in general advances by collecting evidence which verifies old paradigms and refines them. The role of measurement becomes a mere verification of old laws and the description of new ones using ‘irreducible and stubborn facts’ to which the scientist must, by struggle, make his theories conform’ (Kuhn, 1962, p.1-12 -13). However, this advancement may be revolutionary at some point in history when data uncover serious anomalies in a given paradigm, destroying it and giving birth to a new one. This is called a paradigm shift whereby new knowledge creates a serious discrepancy between the data and the paradigm. Kuhn uses the example of measurement and its role in scientific research to illustrate his concept of paradigm shift. Instead of simply corroborating existing laws and creating new ones, the function of measurement should be that of the search for discrepancies.

According to Rasch modellers, the Rasch model with its property of invariance subscribes to criteria of fundamental measurement (Bond & Fox, 2001; Wright, 1997). Given this premise, Andrich (1988) referred to Kuhn’s perspective on the role of measurement in science and argued that the role of the Rasch model in applied science is not mere measurement of a particular trait on a scale. The Rasch model should rather also seek revealing inconsistencies in the data.

According to Andrich (1995), there are at least two approaches to the data-model relationship when constructing instruments. Both approaches invest a lot of effort in constructing test instruments with content validity. And both of them comprise item (test) and person parameters. However, the first approach, as embodied by some models such as 2-parameter

and 3-parameter logistic IRT models, subscribes to the traditional perspective on measurement while the second approach, characterising the Rasch model, is in line with the Kuhnian modern perspective on measurement. The first approach to measurement – attempts to mathematically model the data with very little restriction on the class of models and number of parameters included in the analysis while the second approach imposes *a priori* restrictions on the class of models and parameters that might be used.

In the first approach, a lot of effort is invested to make the model fit the data. When data shows misfit, a new model is generated with more parameters to better account for the data. Misfit is considered as a sign that the model is not functioning appropriately, instead of blindly following a model that does not describe the data observed a new model that better fits the data should be selected. In the second approach, the data must fit the model for meaningful interpretations to be generated. ‘The case for the model does not stand or fall with success or failure in applying it in a particular substantive area (Duncan, 1984, pp. 398-399).’ Misfit is considered as a sign of inconsistency in the data. Generating a new model which accounts better for the data is considered as a way to camouflage discrepancies which should instead be handled by looking for the qualitative sources of misfit.

The main case for the second approach is the property of invariance embodied by the class of models; this property is independent of the data and remains an empirical question. This difference in the approach has led many Rasch proponents to distinguish Rasch from other IRT models by stating that Rasch modelling subscribes to a different paradigm and classified Rasch models under a different theory – the Rasch modelling theory (Andrich, 2004).

In this research, violations of the property of invariance are sought and inconsistency in the functioning of items across the different countries is assessed. In this thesis, the choice of the

Rasch model stems from a pragmatic epistemology rather than the ones portrayed above. The model serves the purpose of the research and seems to be optimal at addressing the research questions. I do not claim to adopt the Kuhnian perspective on the role of the Rasch model as measurement in social sciences. While the theory may appear flawless in terms of its logic and attractive to some, the implications of treating measurement in social sciences as measurement in physical sciences may lead to flawed practices. In the following, I illustrate some of these practices and put forward my stance in relation to measurement in social sciences.

4.3 Measurement in social sciences

In part (4.2.1) of the previous section, it mentioned that due to the property of invariance, comparisons of the difficulty of the same science item administered in two languages, for instance English and French, could be compared using a Francophone person, or even, given the invariance, an Arabic-speaking person. What does it mean when language versions of the same science test or item are of the same difficulty? Is it the *science* part in the item that predicts the difficulty; and hence, it being the same across the languages, the resulting tests would be of the same difficulty? Or does language interact with the difficulty of an item or a test? How can we make sure that items written in different languages are of the same difficulty? Is it a case where a given variable is measured under different conditions just like measuring the volume of a gas under different temperature conditions while keeping the pressure constant? Or is it a case of a confounding variable, difficult to control or eliminate?

In Chapter 2, I argued that language was part of the construct in science assessments including PISA, be it intended by the examiner or not. Hence, the comparison is not made between psychological attributes functioning under different language milieus. In physical science

terms, the *experiment* at hand has a *confounding variable* – the language of administration. Chapters 2 and 3 of this thesis explained that this variable is very complex and cannot be treated as an external condition. Language is inherent in testing and intrinsic to the subject matter. It is a product of the translation-adaptation process and highly reflective of the culture of the examinee. It also shapes ways of thinking of the individual and cannot be separate from him/her. The question at hand amounts to answering the following three questions: (1) what is the nature of the variable of interest? (2) Is this variable measureable? (3) Is measurement in social science possible and what does it mean?

Many of those debates were carried out in the field of quantitative psychology (Borsboom, 2005; Michell, 1997a, b). Educational assessment, being an interdisciplinary field (Baird & Black, 2013), makes use of many of the psychometric techniques and findings in the field of developmental and cognitive psychology. Several of the issues faced in psychology are relevant in educational assessment. In the following, I will be addressing each question in turn in light of Borsboom's (2005) triple perspective approach and relevant literature. I show how answering these questions calls for the ontological stance and may interact with the empirical stance (questions 2 and 3).

In this section, I outline my agreement with the perception of psychological attributes being socially constructed (Wright, 1998). Many of them underlie educational assessments. Until this date, they have not been proven to be quantitative. Whilst scholars like Michell (1997a; 1997b) use this criterion to distinguish scientific measurement from non-scientific measurement (i.e. no measurement at all), I do not object to the use of the term measurement in social sciences. However, given that the variables in physical sciences are essentially different from psychological attributes, measurement in these disciplines cannot be considered

of the same nature. I argue that the operations valid in physical sciences are not necessarily always useful in applied psychology. Instead, many operational measurement methods, not deemed to be ‘scientific measurement’ can be effective in yielding valuable knowledge in psychology and educational assessment.

4.3.1 The nature of the variable – the ontological stance

In Chapters 2 and 3, I argued that in science testing, language is an intrinsic part of the science construct. The question at hand is ontological in nature; determining the nature of the variable depends on the researcher’s ontology. In CTT, the construct is the ‘true score’; it is everything but error. This tends to conform to an operationalist perspective. In LTT, the construct is an unobservable psychological attribute, i.e. a latent trait that can only be indirectly measured using valid tests. There is a relation between the observed score and the theoretical attribute but the former is not a synonym of the latter (Borsboom, 2005). This perception, according to Borsboom, conforms to a realist view. Extreme realists believe that the construct refers to a physical structure present in the human brain.

For representationalists, the score refers to a location of a person on a scale. Representationalists are usually constructivists according to Borsboom. They believe that thinking in terms of constructs is a useful way for psychologists to understand the mind and its activity but that constructs are a creation of the researcher’s genius. These ontological distinctions are very important because they help understand the various epistemologies underlying quantitative psychology and explain much of the diversity in the models and methods adopted in this field.

I adopt the representationalists' perspective to some extent. Latent variables are not physical entities present in the human brain; they are the product of the researcher's imagination (Borsboom, 2005; Wright, 1998). They underlie tests which are socially constructed (Baird & Black, 2013). They have the most value when they are operationally defined as their meaning and understanding changes over time for a variety of reasons. Thinking in terms of constructs makes more tangible the human endeavour into understanding the complexity of the mind (Wright, 1998).

Quantitative psychology mainly aims to measure the variable in question. However, much controversy was manifested around what can be considered as scientifically measureable. The following is an account of the debate and my position in this regard.

4.3.2 Is the variable measureable?

Michell (1997a) claims that psychological attributes are only measureable if they were empirically proven to be quantitative. This is in congruence with Hölder's (1901) theorem: continuous quantities are measureable because they are characterised by a specific internal structure, the additive structure. Michell (1997a) summarises Hölder's axioms into five main conditions to prove that an attribute is quantifiable. Below is a succinct account of Michell's summary:

Given a continuous quantity Q ,

1. any two magnitudes a and b of Q are either:
 - Identical, in which case $a = b$
 - Or different, in which case there must be a third magnitude c in Q such that $a = b + c$ or $b = a + c$

2. for any magnitudes a and b in Q , $a + b = b + a$
3. for any magnitudes a , b and c in Q , $a + (b + c) = (a + b) + c$
4. for any magnitudes a and b in Q such that $a > b$, there exists a magnitude c in Q such that $a > c > b$; and
5. given any two mutually exclusive sets of magnitudes A and B , where A is an ‘upper’ set of magnitudes and B is a ‘lower’ set of magnitudes, if $a \in A$ and $b \in B$ and each magnitude $a_i \in A$ is greater than any magnitude $b_i \in B$, there must exist a magnitude c such that $c < a_i$ and $c > b_i$.

In 1940, the Ferguson Committee, established by the British Association for the Advancement of Science, adopted Hölder’s (1901) conditions for proving that an attribute is quantifiable. Later, the work on the theory of conjoint measurement by Luce and Tukey (1964) and its subsequent developments (Luce et al., 1990) extended Hölder’s work beyond additive systems to empirically test the hypothesis that a given attribute is quantifiable. According to the theory of conjoint measurement proposed by Luce and Tukey (1964), discrete attributes could be established as continuous variables through building special relations amongst the attributes and hence the variables could be proved as quantifiable. Although initial work in proving the quantifiability of an attribute was carried out on continuous variables, the theory of conjoint measurement allows the quantifiability of categorical variables which do not obey Hölder’s strict axioms. For instance, measurement in psychophysics often makes use of categorical data yet the Weberian tradition links the psychological attribute to a physical scale (Morgan, 1997).

Michell (1997a; 1997b) instigated a polemic against quantitative psychologists, claiming that researchers in this field overlooked the scientific endeavour of measurement— i.e. that of proving that a psychological construct is measurable and hence quantitative. Instead, they concentrated on the instrumental task – that of seeking methods to measure psychological

attributes – despite the efforts of the Ferguson Committee to emphasise Hölder's (1901) conditions. Proving empirically that a variable is quantifiable might be possible in physical sciences but what if some attributes in psychology or education are only accessible through human judgement and response to stimuli (Laming, 1997)? Do researchers refrain from carrying out further investigations for lack of evidence that the variable is quantifiable?

From my perspective, research should not be limited by the inability of proving the quantifiability of an attribute. Michell (1997b) adheres to this position by admitting in his response to Laming (1997) that significant discoveries can be made without necessarily being founded on what he believes to be psychophysical measurement. However, I do agree it is necessary for researchers to determine the nature of the variable they are dealing with. If the variable is quantitative, then fundamental measurement is possible. If the quantitative nature cannot be proven, then the methods and the numerical representations adopted should be consistent with the nature of the variable in order for subsequent interpretations to be plausible.

The next section outlines the meaning of measurement from various perspectives and discusses the implications of confusing scientific measurement with measurement in social sciences. Comparing measurement in social sciences to measurement in physical sciences is at the risk of systematically extending the operations carried out on physical quantities to psychological attributes which, to this date, have not been proven to be quantifiable (Michell, 1997a).

4.3.3 Is measurement in social sciences possible and what does it mean?

Hölder (1901) defines scientific measurement as the enterprise of estimating or discovering ratios of a magnitude of a quantity to a unit of the same quantity. From this perspective, scientific measurement is empirical and can only occur on quantitative objects (Michell, 1997a). Kline (1997) claims that psychometricians may never achieve fundamental measurement; in which case, psychology may never be considered as a science.

Does the possibility of scientific measurement in any discipline, and in psychology in particular, make it a science or is it the value of the findings and knowledge generated empirical using rigorous research methods? For many psychologists, there seems to be what Gould (1996) labels as the ‘physics envy’ or the need to establish psychometrics as a science (Michell, 1997a) by establishing numerical representations in psychology. For other psychologists such as Laming (1997), scholars are rather interested in understanding psychological attributes by rank ordering categorical data and are not aspiring for fundamental measurement. This should not make psychology less of a science.

Morgan (1997) and Laming (1997) explain the importance and usefulness of ordering categorical data in psychology. Kline (1997) describes the valuable contribution of non-scientific measurement in psychology such as CTT and operational measurement. According to him, some type of measure is better than none, as these pave the way to further investigations. Intelligence has not yet been proven to be a quantitative attribute. Intelligence is determined until today by what intelligence tests measure (Laming, 1997). Operational measurement carried out in this area generated *operational measures* which proved to be useful predictors and very helpful for theory development (Kline, 1997). Would these findings be demoted to the level of mere incidental relationships?

Michell (1997a) blamed psychologists for adopting Steven's (1959) definition of measurement as 'the assignment of numerals to objects or events according to rule'. While I agree with his objection to having this rule unspecified, making it dependent on the researcher's choice not on the relationships related to the variable revealed empirically (Luce, 1997), I do appreciate the levels of representation scales suggested by Stevens (1959), which are very useful in ordering psychological and educational attributes. Why should measurement be restricted to ratio-scale measurement?

The obsession of having psychology prove itself as a science by performing measurement is comparable to the need to make educational assessment look like a science by virtue of psychometric models. Some prescriptive psychometric models such as the Rasch model subscribe to fundamental measurement (Wright, 1997; Bond & Fox, 2001). However, are the underlying psychological attributes in this context, i.e. certain abilities in science, comparable in the same way physical quantities such as length or mass are comparable on the same scale in physical sciences. In this thesis, similar questions are raised. Psychologists have not yet proven that abilities in social sciences are quantifiable in Hölder's (1901) strict sense of additivity structure or in the broader sense of Luce and Tukey's (1964) conjoint measurement. This means that operations in physical sciences cannot be directly applied to psychology, or else, we end up with comparisons that in reality may mean very little, such as equating scores on two tests very different in nature, for instance, physical sciences and dance or comparing the difficulty of the same item written in different languages using a person from an even different language group. I do not object to the use of the term 'measurement' in psychology as does Michell (1997a; 1997b), yet I concede that there are implications of using this term without clarifying its nuance in social sciences. Variables in social sciences are more complex

and less controllable than in physical sciences. Applying the same operations, valid in physical sciences, in social sciences might lead to controversial results less capable of painting an accurate picture of reality.

In the PISA context, the various scores produced by the tests and questionnaires could be helpful in understanding how educational systems operate. From the Rasch modelling perspective, those scores refer to a location on a scale with an underlying trait. The scientific abilities have not been proven to be quantifiable yet. Hence, according to Michell (1997a), we cannot talk about scientific measurement. However, if we agree to perceive it as a form of social science measurement, various questions can be raised. What does PISA measure in reality? Are comparisons across countries sound despite the distinct populations, cultures, languages, educational systems, curricula, etc.? Is the science ability construct the same across languages? Does social measurement in educational assessment lead to progress in the field? These questions are addressed in the Discussion chapter. The last section in this chapter presents the requirements of the Rasch model and relates some of the criticism in the literature to the way the model is used in this research.

4.4 Requirements of Rasch modelling

A consequence of invariance is that the total score is a sufficient statistic (Andersen, 1977); all the information related to the ability of a person on a given trait is summarised by the total score. This property is shared with CTT but in Rasch modelling, this property is not a mere interpretation of the definition but an outcome of the interaction between a person and an item. In Rasch modelling, all persons with the same total score have the same person estimate.

In addition to invariance, several other requirements must be fulfilled for an instrument to be compatible with fundamental measurement in physics (Andrich, 1988): conjoint additivity, unidimensionality, symmetry and local independence. The Rasch model requires that the instrument used must conform to the theory of conjoint additivity probabilistically, stating that measurement is a ratio of the size of one entity to the size of another defined as a unit (Brodgen, 1977). The measurement instrument must be unidimensional; i.e. the instrument measures a single latent variable (e.g. ability in science). Moreover, the instrument must be designed in a way where the latent construct and single parameter considered are symmetrical.

For instance in educational assessment, the higher the candidate's ability, the more they are able to solve harder items. The test should also be devised in a way where responses to items are independent; i.e. answering one item is not in any way dependent on successfully answering another item. This is often referred to as response dependence to be distinguished from trait dependence which relates to multidimensionality and the violation of the unidimensionality requirement (Marais & Andrich, 2008b). Similar criteria for measurement in social sciences have been earlier articulated by Thurstone in the 1920s and Guttman in the 1940s. These, like Rasch modelling, are consistent with the Kuhnian approach on the role of measurement (Andrich, 2004). According to Andrich, when the data fits the model, it is shorthand for saying that the instrument is invariant across the different groups, is unidimensional, symmetrical and local independence is respected. All these requirements should be empirically tested in datasets. If the data do not fit the model and requirements are not met, the Rasch model can only point to the inconsistency and guide the researcher in the right direction of where to carry out in-depth substantive research.

The Rasch approach has been heavily attacked in the late seventies (Goldstein, 1979) and its use was dropped in the analysis of national examinations in the UK (Panayides et al. 2010). Much of the criticism was related to the model's requirements which many (Goldstein, 1979; Traub, 1983) considered as assumptions. For instance, Traub (1983) states that assumptions like unidimensionality, same discrimination and absence of random guessing, 'fly in the face of common sense and a wealth of empirical evidence accumulated over the last 80 years' (p.64). Proponents of Rasch responded to this criticism by insisting that Rasch does not rely on 'assumptions'; it rather imposes *a priori* requirements on the measuring instrument. Andrich (1989) uses the historical background of the development of Rasch modelling and suggests that the confusion between 'assumptions' and 'requirements' stems from Likert's misunderstanding of the features of Thurstone's scale and his 'law of comparative judgment'. Imposing requirements on the instrument goes hand in hand with ensuring invariance and constructing a suitable measurement tool. The following discussion restricts the critiques of Rasch to the main requirements: the sufficient statistic, the number of parameters, and the unidimensionality of the construct.

4.4.1 The sufficient statistic

In Rasch modelling, the total score is a sufficient statistic and all persons with the same total score have the same person estimate (Andersen, 1977). This property has instigated criticism. People with the same score do not necessarily get the same questions right. To illustrate this point, consider two persons X and Y having the same total score of 5 after sitting the same items. Imagine X reaches this score by getting 5 easy items correct while Y reaches the same score by getting 5 difficult items correct. From the Rasch perspective, X and Y receive the

same person parameter, given their equal total score, whereas some would argue that Y should receive more credit for getting the harder items correct.

Rasch proponents reject this argument for two main reasons. First, because Y has failed on the easier items and for matters of consistency, Y should be penalised severely for getting easy items wrong. The second reason relates to the probabilistic nature of the model. The person parameters are described as a probability function, accurate at a larger sample level. The same total score does not mean qualitatively the same thing at the level of the individual. For the purposes of the study, this argument seems compelling as the focus is not on individual person estimates but rather on item parameters and whether items making up an instrument are invariant across language groups.

4.4.2 The number of parameters

The Rasch model has been criticised for being overly simple by having only one parameter in addition to the person parameter, the item parameter. Other LTT models with 2 or 3 parameters are more likely to show better fit and this might be perhaps useful for specific educational purposes. For instance, Goldstein (1979) argues that in educational assessment, some constructs are set by the curriculum and aligned with specific educational goals. Examiners may purposefully include some items, which may violate requirements of the Rasch model, but with specific curricular objectives in mind. In this case, using 2- and 3-parameter models may be more sensible as the aim is not measurement in the fundamental sense of the term but rather statistical modelling and description of the data. PISA tests are not curriculum based and their purpose operates at the macro level (See Chapter 5 for relevant discussion). Comparing countries fairly only occurs by measuring performances of their

respective samples against the same standard. Diagnosing whether the test used is invariant across all countries (or in this study, across 3 countries of interest) is essential to lay solid grounds for any comparisons made. For this purpose, the Rasch model seems the optimal choice.

4.4.3 Unidimensionality of the construct

Another major criticism is that of unidimensionality, i.e. measuring a single trait. Users of Rasch modelling insist on the importance of this requirement; otherwise, bias may result in item and person estimates (Reckase, 1979). Andrich (2006) mentions the thickness of a variable. Any item/test can be unidimensional at a certain degree of precision just like any item/test can be multidimensional at a certain degree of precision. Smith (2002) argues that unidimensionality is not a ‘black and white’ judgment; but rather a matter of degree and that the challenge is to determine the point at which this degree affects person and item parameters.

Andrich (2014) provides the example of the assessment of physical sciences. One can develop an assessment with various items measuring a specific *thin* construct relating to the topic of electricity, or design a test which measures a *thick* construct of ability in physics which includes questions covering various topics in physics (heat, light, sound, mechanics, electricity, magnetism). According to Andrich (2006; 2014), the latter case does not portray an example of multidimensionality. It is a case where the variable is thick; i.e. it encompasses various ‘sub-abilities’ which constitute thinner variables. The thickness of the dimension is determined by the purpose of the test.

Stenner, Burdick and Stone (2008; 2009) distinguished two types of variables: index variables (also called formative variables) and causal variables (also referred to as reflective variables). Stenner et al. use the example of socioeconomic status (SES) index to illustrate what index variables are. They are thick variables constituted by aggregating indicators of the underlying concept of interest. In the case of SES, the indicators are education, occupation, income, etc. and items referring to these are not exchangeable. As an example of causal variables, Stenner et al. give the example of reading proficiency and claim that items in a reading proficiency test are exchangeable because they measure somehow a person's ability in reading; a single concept.

Andrich (2014) used this distinction to argue that the distinction between index and causal variables is a matter of degree and that thicker higher order variables are index variables while thinner lower order variables are causal variables. Hence, the example of the thick construct 'ability in physics' described earlier is an example of an index variable. It is not uncommon to find students who do well in mechanics but struggle with electricity concepts. This means that items are not exchangeable. However, each of these thinner variables can be considered as a causal variable. Andrich argues that although the indicators which make up the thick variable in this case are distinct, at a deeper level of knowledge, they define a single thick physics variable. Stenner et al. (2008; 2009) highlighted that even if the data fit the Rasch model and thus the requirement of unidimensionality was not violated, this does not allow the distinction of the nature of the variable: is it an index variable or a causal variable. Additional empirical evidence is required.

In Chapter 2, I argued that language is an intrinsic part of the scientific literacy construct in PISA science 2006 tests. In this case, scientific literacy cannot be considered as a thin causal

variable as its constituents are quite distinct. The scientific literacy construct is a thick index variable consisting of the scientific knowledge, thought processes and communication competencies. However, the link between scientific thinking and language outlined in Chapter 2 makes the two variables inextricable. The distinction between items targeting scientific knowledge and items measuring proficiency in scientific language becomes harder to define.

Goldstein, Bonnet and Rocher (2007) analysed data of PISA 2000 reading test of England and France using multi-level modelling and structural equation techniques. Results suggest that the test contained at least two dimensions. Moreover, Kreiner and Christensen's (2014) recent analyses on data of the PISA 2006 reading test provided evidence of differential item functioning and hence of multidimensionality based on Marais and Andrich (2008). No such analysis on PISA science items has been published in the literature yet. In Chapter 6, I analyse PISA science 2006 data using the Rasch model to examine whether the data fit the model. I carry out a principle component analysis to investigate whether scientific literacy is a single thick index variable given an acceptable degree of precision or whether it is a multidimensional variable. This point will be discussed again in Chapter 6.

This research investigates whether data violates the requirement of unidimensionality using principal component analysis (Chapter 6) and carries out a DIF analysis to examine how items behave in the different countries (Chapter 7).

CONCLUSION

PISA uses the Rasch model to calibrate items because it is a highly suitable model for making comparisons across countries on a unidimensional scale (OECD, 2009a). However, proponents of Rasch modelling make strong claims about the model's subscription to

fundamental measurement in social sciences and hence its ability to uncover inconsistencies in the data. These claims have great implications on the way test results are handled and interpreted. In this chapter, I take a distance from these strong claims and justify the use of Rasch modelling to analyse PISA science 2006 data purely on pragmatic grounds. It is the model used so far by PISA since its inception in 2000. Besides, it is a simple yet elegant model that places student abilities and item difficulties on the same unidimensional scale making it very suitable for international comparisons.

Tackling the question of fundamental measurement in social sciences cannot be carried out without stumbling upon elemental ontological questions such as whether the variable in question is quantifiable and hence whether it is measureable. These questions are particularly relevant to my research question as they raise concerns about the nature of the construct in question. In my view, scientific abilities do not refer to entities present independent of the researcher's imagination. They are useful artefacts to better understand and measure the human mind. While I agree with the necessity of identifying the nature of the constructs in terms of their quantifiability, I do not believe education and psychology are less worthy if scientific measurement could not be carried out. Concern is raised regarding the lack of nuance in the meaning of measurement in social sciences as many of the operations valid in physical sciences could lead to practices of little use and poor interpretations in social sciences.

This research investigates the extent to which language versions of the same science test are of comparable difficulty and demands. If the data fit the Rasch model, this means that the instrument shows invariance across language groups, unidimensionality, symmetry and local dependence. These properties are true to an acceptable degree of precision defined by the

purpose of the test. Of particular interest is the requirement of unidimensionality. I hypothesised in this chapter that scientific literacy can be considered as a thick index variable. This hypothesis will be tested in Chapter 6. Data of PISA science 2006 from three countries (UK, France and Jordan) is analysed using the Rasch model. Fit of the data to the model is explored and a principle component analysis is used to investigate whether the unidimensionality property is violated. The next chapter however describes PISA science assessments. It provides an overview of their development, design, format and data analysis.

CHAPTER 5

PISA Tests and Sample Description

‘Å vere midt i smørauget’
‘to be in the middle of the butter melting in the porridge’
(Literal translation from Nynorsk)

INTRODUCTION

PISA is one of the most influential international large-scale assessments administered to 15-year-olds by the OECD in as many as 75 countries to date. PISA aims to evaluate educational systems worldwide and make international comparisons possible. The first administration of PISA took place in 2000 with 32 participating countries. The fast growth of interest in the study led to its extension to another 11 countries in 2001 (OECD, 2003). The study has been carried out every three years since 2000 with the last cycle taking place in 2012, with 65 countries taking part. The assessment focuses on three domains – reading, mathematics and science – with one of them being a major domain every year of assessment. PISA data has been considered by many countries as providing valuable empirical evidence for their educational system and has instigated policy change in various countries such as Germany (Ertl, 2006) and Japan (Schleicher, 2009; Takayama, 2008). Unlike other large-scale assessments such as TIMSS and PIRLS, PISA claims that its assessment is not curriculum-based. The test measures students’ ability to transfer knowledge and use it in real-life situations and their capacity to be active members of their societies at the end of compulsory education.

Reading was the major subject in the first cycle in 2000. The last administration of PISA took place in 2012 with mathematics being the major subject. This research investigates the

differential impact of language on the difficulty of science tests; hence, data of PISA science 2006 test was considered as it is so far the only administration of PISA with science being the major domain. Science will again be the focal domain in the next cycle; that is, in PISA 2015. In PISA 2006, fifty-eight countries participated; a total of 87 national versions in 44 languages have been developed (OECD, 2009b). In this research, data of three countries – UK, France and Jordan – administering PISA in different languages, respectively English, French and Arabic were analysed.

The first section of this chapter describes the format and design of the PISA test. The second section describes the scientific literacy framework adopted in PISA 2006 science tests. The third section outlines the item development process. The fourth section summarises the guidelines adopted for test translation and adaptation. The fifth section outlines the sampling design. The sixth section describes the models used for data analysis by the OECD, the item calibration and discusses the way results are reported. The last section provides an overview of the three countries considered in this research. This background information about the construction of the tests and scales is necessary to understand the data analysed in this thesis.

5.1 The PISA science test

To avoid cultural bias due to item familiarity, PISA uses a variety of item types and ensures that time is not a factor threatening the reliability of the test. The format of the test is described in the first part of this section. The second part describes the incomplete design adopted by PISA to ensure coverage of the construct and better validity and reliability estimation by rotating booklets consisting of different subsets of questions. PISA science cognitive items are developed following a framework based on research in science education

promoting scientific literacy. The framework was described in Chapter 2 and will be briefly reviewed in the third part. The laborious process of developing PISA cognitive items is detailed in the fourth part of this section. The last part lays out the rigorous techniques used in translating and adapting PISA instruments and materials.

5.1.1 PISA test format

PISA science items are organised into units. Each unit consists of up to four items related to a common stimulus. The stimulus could be a text, a table, a graph, a diagram or a combination of those. Administered science units comprised two types of items: cognitive items and attitudinal items. Cognitive items assessed students' scientific competencies and knowledge and attitudinal items, introduced the first time in PISA 2006, assessed students' attitudes towards science. PISA 2006 consisted of 37 science units, comprising a total of 108¹³ cognitive items and 31 attitudinal items making up 210 minutes of testing time for science consistent with the amount of time allocated to mathematics, the major domain in 2003. This research project was concerned with the 'cognitive items' and did not analyse 'attitudinal items' as the latter form another scale of measurement.

Twenty-two items amongst the 108 cognitive items used in the main study were 'linking items' previously used in PISA 2003 study. The remaining 86 were selected from a large pool of newly-developed items after extensive field tests conducted in all countries in 2005. PISA 2006 included 8 units (31 items) of Reading used in PISA 2003 making up around 60 minutes

¹³ The international dataset on which all analyses will be based (Chapters 6-8) consists of 103 items. Five items were deleted by the consortium's manager at ACER because of the poor quality of the student information they generated or because they targeted misconceptions (OECD, 2009b).

of testing time and 31 units (48 items) of mathematics selected amongst the ones used in 2003 and representing approximately 120 minutes of testing time.

Science cognitive items consisted of multiple choice questions (standard or complex forms with four distractors with one or several correct choices, true/false or correct/incorrect statements, Yes/ No questions), short closed-constructed responses where students were asked to provide a numeric response, or a word or a short phrase, and open- (extended) constructed responses which require extended written answers.

5.1.2 PISA test design

PISA 2006 items were allocated to thirteen item clusters, each representing approximately 30 minutes of testing time. The thirteen clusters comprised seven science clusters, two reading clusters and four mathematics clusters. PISA test was administered following a balanced incomplete block design where clusters were rotated in a specific way to form thirteen test booklets as shown in Table 5.1 below (reproduced from OECD, 2009b, p.39).

Table 5.1: Cluster rotation design used to form test booklets for PISA 2006 (OECD, 2009b, p.39)

Booklet	Cluster 1	Cluster 2	Cluster 3	Cluster 4
1	S ₁	S ₂	S ₄	S ₇
2	S ₂	S ₃	M ₃	R ₁
3	S ₃	S ₄	M ₄	M ₁
4	S ₄	M ₃	S ₅	M ₂
5	S ₅	S ₆	S ₇	S ₃
6	S ₆	R ₂	R ₁	S ₄
7	S ₇	R ₁	M ₂	M ₄
8	M ₁	M ₂	S ₂	S ₆
9	M ₂	S ₁	S ₃	R ₂
10	M ₃	M ₄	S ₆	S ₁
11	M ₄	S ₅	R ₂	S ₂
12	R ₁	M ₁	S ₁	S ₅
13	R ₂	S ₇	M ₁	M ₃

S: Science cluster; M: Maths cluster; R: Reading cluster; Digit subscripts distinguish clusters within a domain

Each booklet comprised 4 clusters making up a total of 2 hours of testing. In Table 5.1, science clusters were designated by the initial S, maths clusters by the initial M and the reading clusters by the initial R. A digit subscript distinguished the various clusters within a domain (S₁, S₂, M₃, R₂, etc.) This incomplete design ensured that each cluster appeared only once in each of the four possible positions within a booklet, that each test item appeared in four of the test booklets and that each pair of clusters appeared in only one booklet. This design is typically characterized by a large amount of structural missing data and involves a great deal of imputation using plausible values (see Section 2d of this chapter).

It is worth noting that this design means that different students attempt different proportions of questions within each domain. For instance, students who had sat Booklet 2 consisting of clusters S₂, S₃, M₃ and R₁, solved more science items than students who had sat Booklet 7

consisting of clusters S7, R1, M2 and M4. An even more interesting and debatable consequence of the design is that some students received ability estimation for domains for which they had not solved a single question. For instance, students who answered questions of ‘Booklet 1’, consisting of clusters S1, S2, S3 and S7, had their ability in reading and mathematics estimated based on their performance in science, not based on any reading or mathematics item. This design has been fiercely criticised by some scholars such as Kreiner and Christensen (2014).

An additional *Une Heure* booklet (UH booklet) was administered with half the number of items in the other booklets with a proportion of 50% for science, 25% for reading and 25% for mathematics to students with special needs. Items were selected from the main study based on their suitability to special needs students.

5.2 PISA 2006 Scientific Literacy Framework

As mentioned in Chapter 2, an assessment framework is developed for each PISA subject domain. The framework defines the scope of the assessment, details the test structure, describes the item format and the preferred distribution of items according to important framework variables and outlines the possibilities for reporting results. The PISA science 2006 framework focused on scientific literacy and was developed through several stages from December 2002 with its final version published in early 2006 following the integration of amendments suggested by field trials. The OECD definition of scientific literacy as well as the scientific framework were presented and thoroughly discussed in Chapter 2. They are briefly recapitulated below for the sake of completeness.

PISA 2006 defines scientific literacy as follows:

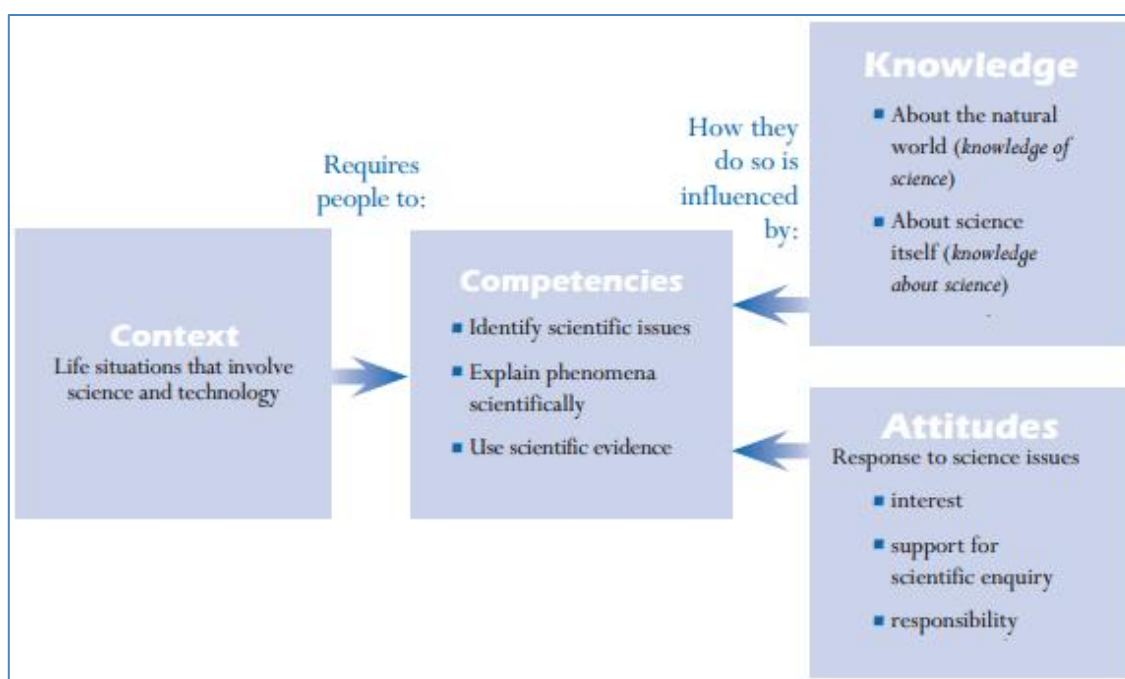
Scientific literacy refers to an individual's:

- Scientific knowledge and use of that knowledge to identify questions, acquire new knowledge, explain scientific phenomena and draw evidence-based conclusions about science-related issues
- Understanding of the characteristic features of science as a form of human knowledge and enquiry
- Awareness of how science and technology shape our material, intellectual, and cultural environments
- Willingness to engage in science-related issues and with the ideas of science, as a reflective citizen

(OECD, 2006, p.23)

Scientific literacy is assessed in relation to scientific knowledge and concepts, scientific processes and situations or context. Figure 5.1 below is a reproduction of Figure 1.1 (OECD, 2006, p.26) which depicts the scientific literacy framework.

Figure 5.1: Framework of PISA science 2006 assessment (OECD, 2006, p.26)



The distribution of the science items, with respect to the major framework categories, is summarised in Tables 5.2, 5.3 and 5.4 reproduced from the technical report (OECD, 2009b, pp.43-44).

Table 5.2: Number and percentage of science cognitive items in terms of item format and competency (OECD, 2009b, p.43)

Item format	Scientific Competency			Total
	Identifying scientific issues	Explaining scientific phenomena	Using scientific evidence	
Multiple choice	9	22	7	38 (35%)
Complex multiple choice	10	11	8	29 (27%)
Closed-constructed response	0	4	1	5 (5%)
Open-constructed response	5	16	15	36 (33%)
Total	24 (22%)	53 (49%)	31 (29%)	108 (100%)

Based on Table 5.2, around a fifth (22%) of PISA science 2006 cognitive items measure the first competency ‘Identifying scientific issues’, almost a half (49%) measure the second competency ‘Explaining scientific phenomena’ and less than a third (29%) of the items measure the third competency ‘Using scientific evidence’. It is clear that the PISA science items are mainly multiple choice items (62% multiple choice items). However, participants are required to formulate explanations and justifications in a large proportion of questions; a third of the items require open-constructed response. Only 5% of questions are closed-constructed questions.

Table 5.3: Number and percentage of science cognitive items in terms of item format and knowledge type (OECD, 2009b, p.44)

Item format	Knowledge of science	Knowledge about science	Total
Multiple choice	24	14	38 (35%)
Complex multiple choice	15	14	29 (27%)
Closed-constructed response	4	1	5 (5%)
Open-constructed response	19	17	36 (33%)
Total	62 (57%)	46 (43%)	108 (100%)

Table 5.3 suggests a reasonable balance between the two types of knowledge involved in science literacy: knowledge *of* science (57%) and knowledge *about* science (43%). This indicates a fair coverage of the scientific literacy construct.

Table 5.4: Number and percentage of science cognitive items in terms of knowledge category and competency (OECD, 2009b, p.44)

Item format	Scientific Competency			Total
	Identifying scientific issues	Explaining scientific phenomena	Using scientific evidence	
Physical systems		15	2	17 (13%)
Living systems		24	1	25 (23%)
Earth & space systems		12	0	12 (11%)
Technology systems		2	6	8 (7%)
Scientific enquiry	24		1	25 (23%)
Scientific explanation	0		21	21 (19%)
Total	24 (22%)	53 (49%)	31 (29%)	108 (100%)

Science cognitive items focused mainly on three knowledge categories: living systems (23%), scientific enquiry (23%) and scientific explanation (19%). Physical systems (13%), earth and space systems (11%) and technology systems (7%) were less emphasized.

5.3 Development of cognitive items

Several research centres were involved in the development of PISA science 2006 items to ensure a wider cross-cultural and cross-national diversity: ACER (Australian Council for Educational Research, Australia), CITO (Central Institute for Test Development, The Netherlands), ILS (Department of Teacher Education and School Research, University of Oslo, Norway), IPN (Leibniz Institute for Science Education, University of Kiel, Germany) and NIER (National Institute for Educational Policy, Japan). The consortium's manager of test and framework development of science at ACER oversaw the work of the development teams. This section describes briefly the development of cognitive items.

The development of cognitive items consisted of several stages following a cyclical process where items would be reviewed, piloted and amended based on feedback then trialled again and improved further. This section provides a brief summary of the process and only presents the broad stages of development. Chapter 2 of the technical report (OECD, 2009b) provides all the details of the process.

In a first phase, test units were developed by development teams in the different centres in the local language. Later, items went through cognitive laboratory activities which included local item panelling for extensive item and coding evaluation, cognitive interviews and '*think-aloud*' methods for readability examination, and local item testing and piloting for preliminary empirical data gathering. Item functioning and difficulty were evaluated and authentic student responses were compiled. This triggered another round of refinement of item wording and coding guides.

In a second phase, items underwent international item panelling where units were reviewed and scrutinised by other development teams. Later, units were field tested in different schools, in Australia only, due to translation and timeline constraints. Many items were revised in a minor fashion. If items required extensive revision, they were either discarded or went through another round of piloting and field testing.

Participating countries were asked to provide units of science cognitive items following detailed guidelines to ensure maximal cultural and contextual representation. After careful scrutiny of proposed national items by development centres, units were circulated to national centres for review. National centres rated items and provided feedback on the extent to which items were relevant and interesting to 15-year-olds in the particular country, the degree to which they were related to material covered in the national science curriculum and how well they were related to the scientific literacy framework. Later, these item ratings led to either revising or discarding items.

At a later stage, a forum was organised in July 2004 in Warsaw where experts from all participating countries were invited to partake in a detailed international item review. Subsequently, all test instruments were developed into English and French. These became the source versions which countries adapted and translated to form their national versions. The English and French source versions were developed in a parallel manner in a way to identify and address issues arising from content and wording immediately and to limit cultural bias in these versions. More details on the development of the source versions and their adaptation and translation process are described in Section 1e of this chapter.

In a fourth phase, cognitive items were selected for the field trial after a lengthy process where items were further reviewed; some were improved while others were weeded out. The final selection of units was informed by the reports produced during the national item review phase. These units were sent to the national centres where rigorous adaptation and translation into different languages took place and experts in each country were asked to verify whether the final test forms were equivalent. Later, items were field tested with 15-year old students in different schools in participating countries. Extensive analyses of the field trial data were carried out. Item responses were extensively analysed for item fit, item discrimination, item difficulty, distractor analysis, mean ability, gender and country bias, etc. These analyses resulted in further weeding of items and refinement of others. An additional sequence of national reviews was conducted using the outcomes of the field tests of items. A timing study was conducted to determine the number of items to be included in the main study booklets. Clusters of items were formed and booklets were assembled based on the cluster rotation design provided earlier in Table 5.1.

PISA responses are coded by trained individuals and not directly scored. That is, coders are supposed to identify student responses with the various types of possible responses indicated in coding manuals and then attach a specific code. Different codes do not necessarily result in different scores. International coder training sessions for all PISA domains were conducted. These training sessions helped further refinement of coding guides. Subsequently, national centres organised local coding trainings. A coder query service was established online to assist the coding of scripts in each country. This process was repeated for the main study and the coder query service was maintained to constantly provide coding advice.

The main study consisted of 37 units comprising 108 cognitive items; eight of these units were kept intact as these were link units administered in 2003. That is, these link units ensured comparability of scores in science between the 2003 and 2006 cycles by enabling science scores of the two cycles to be put on a common measurement scale. Item responses were extensively analysed. Five items were removed from the international datasets because they either generated poor student information or focused on misconceptions. Hence, a total of 103 cognitive items formed the science cognitive scale.

As described above, PISA invests tremendous efforts in developing items that are not culturally biased. Various rounds of revisions and refinement are carried out using several techniques (item panelling, cognitive interviews, piloting, etc.). Participating countries are encouraged to suggest units of items in an attempt to decentre the test as much as possible. This is one of the recommendations suggested in Chapter 3 (Section 3.5.2.c) to make tests easier to translate and adapt. The process is thorough with only five items deleted from international datasets. The intense training of coders and the establishment of an online coder query service ensure that coding is consistent within and across countries.

5.4 Translation and adaption of PISA 2006

Chapter 3 discussed the literature on test translation and adaptation. It examined international guidelines for the translation and the adaptation of tests, the various methodologies adopted and a number of issues associated with the validity and reliability of cross-cultural comparisons. This section is limited to the description of the guidelines and processes adopted by PISA for creating language versions of same surveys.

PISA implements rigorous methods and verification procedures for translating and adapting tests as its aim extends beyond ranking countries and students. One of PISA's main purposes is to provide descriptive scales of students' literacy comparable across countries. Poor translation or adaptation of instruments and significant bias would lead to unreliable and invalid comparisons. To minimize cultural bias, PISA carries out a double translation from two source versions (English and French) of its instruments.

The section below describes procedures adopted by PISA including the development of two source versions, the double translation process and the quality control procedures put in place to verify the quality of the translation and adaptation of tests.

5.4.1 Development of English and French source versions

As discussed earlier, a number of items used in PISA 2006 were link items kept intact from PISA 2003. The new items used in PISA 2006 were either developed by the previously mentioned test development teams (i.e. ACER, CITO, ILS, IPN and NIER) or proposed by some of the participating countries. Development teams were provided with guidance on how to anticipate potential translation and adaptation problems such as using particular idioms or cultural expressions. These guidelines were informed by experience accumulated from previous PISA cycles. PISA encouraged items to be developed in the country's national language; however, most items were submitted in English. The number of items submitted in English or in other national languages was not mentioned in any of the technical reports (see OECD, 2009b, p.33). The French source version was developed through double translation and reconciliation with the English version of the test instruments and materials into French. A French domain expert as well as a native professional proof-reader reviewed the final

French source version to verify the adequacy of the scientific language and the linguistic quality. Later, a bilingual expert verified the equivalence of the English and French source versions using specific criteria and defined procedures.

5.4.2 Double translation from two source languages

As described in Chapter 3, there are various methods to translate and adapt tests into different languages and cultural settings. Direct translation was highly discouraged in the literature for reasons listed in that chapter and back-translation was advocated as it has proved to result in a better quality of translation. PISA adopted a particular form of test translation/adaptation: double translation of two source versions into the target language followed by a reconciliation procedure. This method consists of independently translating the source versions into two separate versions in the target language and then merging the two versions into a final version in the target language. This method ensures that idiosyncrasies embedded within a particular language are detected during the adaptation process and will thus have less impact on the development of target versions.

Eighty-seven national versions have been developed in 44 languages. National versions, except those used with minorities making less than 5% of a target population or those borrowed without further adaptations, went through an elaborate international verification process involving various rounds of revisions and checks of the item units, questionnaires and manuals as well as coding guides. In this process, a trained team of expert reviewers examined each national version against the English and the French source versions for equivalence.

Several analyses were carried out on data of the main study to examine the quality of the national version. Different statistics were generated at country level and at item level. At country level, an item-by-country coefficient was computed to investigate the quality of the national versions. The mean absolute deviation (MAD) and the root mean squared error (RMSE) as well as a chi-square statistic were used as indices summarizing item-by-country interaction of each language within a country. Results are detailed in Chapter 5 of the PISA 2006 technical report (OECD, 2009b). At item level, item format or item focus did not seem to have any impact on the item-by-country interaction average; however, the assessed competency seemed to be highly associated with the item-by-country interaction average with a mean interaction ranging between 0.33 and 0.36 for the different three competencies.

Twenty-eight out of 179 PISA 2006 items were eliminated for various reasons, mainly because of poor translation, poor printing quality or layout issues. This is quite a high number of ‘defective’ items considering the extensive revision that items go through in the development and adaptation processes.

5.5 Sampling design and item calibration

PISA’s target population consisted of 15-year-old students enrolled in educational institutions within participating countries. A sampling frame was developed based on specific criteria for eligibility or exclusion of students and schools which ensured optimal representation of countries, limited estimate bias and yielded reliable results. Schools in the sampling frame were categorised according to stratifying variables, such as school size, school type, region, urbanicity, etc. (See STRATUM variable in the data file *INT_Stu06_SAS_Dec07.txt* and the associated SAS file).

PISA adopted a two-stage sampling approach for all participating countries except for the Russian Federation where a three-stage sampling design was adopted. Schools (variable identified as SCHOOLID in the data file) were considered the Primary Sampling Unit (PSUs) in PISA. In the first stage of the sampling, PSUs were selected within each country from the pre-defined target population with probabilities of selection that were proportional to the school size. This way, in smaller schools which have a lower probability of being selected, students will have a higher within-school probability of selection and vice versa. A minimum of 150 schools was normally required per country. In the second stage, 15-year-old students were selected from participating schools following stratified random sampling. Typically, a minimum of 35 students was selected from every school; and a minimum of 4,500 assessed students was required for a country to participate in the study.

Some units in the sample are more important than others and have to contribute more than others for any population estimates. School and student weights were therefore used in the analysis of PISA data to ensure that each sampled student represents the correct number of students in the PISA target population. The weight variable is symbolized by W_FSTUWT in the data file.

5.6 Data Analysis

Data analysis of PISA makes use of a generalised form of the Rasch model and a population model. This section describes these models consecutively.

5.6.1 The mixed coefficient logit model

PISA applies a generalized form of the Rasch model – the mixed coefficient logit model – where items are described by a vector $\boldsymbol{\xi}$, with the latent variable, $\boldsymbol{\theta}$, describing student abilities in a particular domain, being a random effect.

For I items, indexed $i = 1, \dots, I$ and with each item having $K_i + 1$ response categories indexed $k = 0, 1, \dots, K_i$, the random variable can be expressed as $\mathbf{X}_i = (X_{i1}, X_{i2}, \dots, X_{iK_i})^T$, where

$$X_{ij} = \begin{cases} 1 & \text{if response to item } i \text{ is in category } j \\ 0 & \text{otherwise} \end{cases} \quad (5.1)$$

to indicate the $K_i + 1$ possible responses to item i .

The zero category is arbitrarily used as a reference category for model identification and is expressed as a vector of zeroes. The $\mathbf{X}_i$ can also be collected together into the single response vector $\mathbf{X}^T = (X_1^T, X_2^T, \dots, X_I^T)$ with the particular instances of each of these random variables denoted as x , x_i and x_{ik} .

A vector $\boldsymbol{\xi}^T = (\xi_1, \xi_2, \dots, \xi_p)$ of p parameters describes items. Response categories of each item are described in the response probability model using linear combinations of the p parameters. D , design vectors $\mathbf{a}_{ij}$, ($i = 1, \dots, I$; $j = 1, \dots, K_i$), each of length p , can be collected into a design matrix $\mathbf{A}^T = (\mathbf{a}_{11}, \mathbf{a}_{12}, \dots, \mathbf{a}_{1K_1}, \mathbf{a}_{21}, \dots, \mathbf{a}_{2K_2}, \dots, \mathbf{a}_{IK_I})$.

The vector $\boldsymbol{\theta} = \theta_1, \theta_2, \dots, \theta_p$, represents an individual's position in the D -dimensional latent space defined by a set of D latent traits underlying individuals' responses.

A scoring function is introduced into the model to specify the score assigned to each possible response category for each item. Hence, the response score b_{ijd} designates the scores of an observed response in category j to item i , on dimension d . Across D dimensions, scores are collected into a column vector $\mathbf{b}_{ik} = (b_{ik1}, b_{ik2}, \dots, b_{ikD})^T$ and into a scoring sub-matrix for item i , $\mathbf{B}_i = (b_{i1}, b_{i2}, \dots, b_{iD})^T$. Scores are finally collected into a scoring matrix for the entire test:

$$\mathbf{B} = (\mathbf{B}_1^T, \mathbf{B}_2^T, \dots, \mathbf{B}_I^T)^T.$$

The probability of a response in category j of item i is modelled as

$$\Pr(X_{ij} = 1; A, B, \xi | \theta) = \frac{\exp(b_{ij}\theta + \mathbf{a}'_{ij}\xi)}{\sum_{k=1}^{K_i} \exp(b_{ik}\theta + \mathbf{a}'_{ik}\xi)} \quad (5.2)$$

For a response vector, we have:

$$F(x; \xi | \theta) = \Psi(\theta, \xi) \exp [\mathbf{x}'(\mathbf{B}\theta + \mathbf{A}\xi)]; \quad (5.3)$$

with

$$\Psi(\theta, \xi) = \{\sum_{\mathbf{z} \in \Omega} \exp[\mathbf{z}^T(\mathbf{B}\theta + \mathbf{A}\xi)]\}^{-1}; \quad (5.4)$$

where Ω is the set of all possible response vectors.

5.6.2 Population model

The item response model is conditional on the latent variable, θ . Let α symbolise a set of parameters that characterise the distribution of θ . Assuming that students are drawn from a normally distributed population with mean μ and variance σ^2 . That is:

$$\boldsymbol{\theta} = \mu + E \text{ where } E \sim N(0, \sigma^2) \quad (5.5)$$

where the distribution of E is the same as that for $\boldsymbol{\theta}$, but with a mean of zero.

Equation 5.5 is a very simple linear model for variation between students in $\boldsymbol{\theta}$. We extend this model by replacing the mean, μ , with the regression model $\mathbf{Y}_n^T \boldsymbol{\beta}$ (Adams et al., 1997). $\mathbf{Y}_n$ is a vector of fixed and known predictors for student n such as gender, socioeconomic status, country membership. $\boldsymbol{\beta}$ is the corresponding vector of regression coefficients. The population model becomes

$$\boldsymbol{\theta}_n = \mathbf{Y}_n^T \boldsymbol{\beta} + E_n \quad (5.6)$$

where E_n is assumed to be independently and identically normally distributed with mean zero and variance σ^2 so that (5.6) is equivalent to

$$f_{\boldsymbol{\theta}}(\theta_n; \mathbf{Y}_n, \boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma^2} (\theta_n - \mathbf{Y}_n^T \boldsymbol{\beta})^T (\theta_n - \mathbf{Y}_n^T \boldsymbol{\beta})\right], \quad (5.7)$$

a normal distribution with mean $\mathbf{Y}_n^T \boldsymbol{\beta}$ and variance σ^2 .

Equation 5.7 can be further generalized into the multivariate population model below to apply it to the vector-valued $\boldsymbol{\theta}$. The extension results:

$$f_{\boldsymbol{\theta}}(\boldsymbol{\theta}_n; \mathbf{W}_n, \boldsymbol{\gamma}, \Sigma) = (2\pi)^{-\frac{1}{2}} |\Sigma|^{-1/2} \exp\left[-\frac{1}{2} (\boldsymbol{\theta}_n - \boldsymbol{\gamma} \mathbf{W}_n)^T \Sigma^{-1} (\boldsymbol{\theta}_n - \boldsymbol{\gamma} \mathbf{W}_n)\right], \quad (5.8)$$

where $\boldsymbol{\gamma}$ is a $u \times d$ matrix of regression coefficients, Σ is a $d \times d$ variance-covariance matrix, and $\mathbf{W}_n$ is a $u \times 1$ vector of fixed variables where $\mathbf{W}_n$ variables are referred to as conditioning variables.

5.6.3 PISA item calibration

International item parameters were set by applying the mixed coefficient logit model (5.3) in conjunction with the multivariate population model (5.8) to a subsample of students referred to as the international calibration sample. This subsample consisted of 15,000 students formed by the random selection of 500 students from each of the 30 participating OECD countries. The international calibration sample was used to calibrate PISA item parameters (OECD, 2009a; 2009b). It is standard procedure to reduce sample size when examining data fit to the Rasch model as tests of fit are sensitive to sample size. However, it was not clear why PISA chose an extremely small subsample of students per country. Besides, the consideration of an equal number of students for all countries despite the large variation of country size was not justified in any of the technical reports.

It is also unclear if the sampling design (i.e. the strata) was taken into account in the item calibration process. The conditional item response model was adopted alongside the multivariate population model to estimate international item parameters. Five¹⁴ separate one-dimensional models were used.

5.6.3.a *Conditioning variables*

Conditioning variables included direct regressors: booklet, gender, and school (OECD, 2009b, p. 154). Indirect regressors used in the scaling model were defined based on a principle component analysis of all other categorical variables from the student, ICT and parent questionnaires (See Appendix 2 in OECD, 2009b).

¹⁴ The technical report mentions four scales and four one-dimensional models instead of five (OECD, 2009, p.219). This is believed to be a mistake. There are clearly five scales: Reading, Maths, Science, ICT and Interest.

5.6.3.b *Plausible values in PISA*

PISA uses an incomplete design where students sit different subsets of items. Imputation of the missing data is carried out using plausible values (PVs) to create continuous distributions of variables. PVs are random draws from the marginal posterior distribution of the latent variable for each student (Mislevy & Sheehan, 1987). Posterior distributions are distributions mathematically generated around reported values, in this case student scores. The PISA 2006 technical report refers to its 2003 edition for details about procedures of drawing PVs in PISA. These were based on Mislevy (1991) and Mislevy et al. (1992) where PVs are drawn twice; first, they are drawn using parameters from the latent regression model and then they are again drawn from the posterior distribution. The posterior distribution is the distribution of unobserved data conditional on the observed data. In PISA, five PVs are generated for student performance.

In the PISA data analysis manual (2nd edition) (OECD, 2009a), chapter 6 justified the use of PVs over other Rasch measures (e.g. *Expected A Priori* (EAP)). However, nowhere do they disclose the posterior distribution used to draw their PVs. Besides, as noted by Kreiner and Christensen (2014), it is not clear how PISA justifies estimating PVs on a given subject (e.g. reading items) for individuals who have not taken any item in that subject (see section 1.b). Typically, performance on other subjects is used in the estimation process; however, deducing performance in maths based on performance in reading is conceptually weak.

5.6.3.c *Dimensions in science*

Science cognitive items of PISA 2006 were developed based on the framework of scientific literacy discussed in a previous section of this chapter (Section 1c). Dimensions in science

mirrored the three competencies described in the framework: (1) Identifying scientific issues; (2) Explaining phenomena scientifically; (3) Using scientific evidence.

5.6.4 OECD results

PISA publishes detailed results of student performance per country and per domain in various volumes (cf: OECD, 2011, 2013b). Public attention is however focused on just one outcome: the mean OECD scores of countries and country rankings in every PISA domain. PISA results have had a high impact in some countries. Germany for instance experienced a '*PISA shock*' in 2003 because of their relative low ranking on PISA international scale (Ertl, 2006). This shock has led to substantial changes in the educational system in both countries. High ranking countries like Finland witnessed a phenomenon referred to as '*PISA tourism*' where policy makers and educational researchers visited the country to investigate ways of improving their own system by adopting some policies which proved successful in Finland (Sahlberg, 2011). PISA's influence was also witnessed in the surge of number of publications relating to international tests (Domínguez et al., 2012).

Kreiner and Christensen (2014) have strongly criticised the country ranking procedure of PISA in their fierce attack on the use of Rasch model in PISA reading literacy. The authors claimed that country rankings did not stand on 'robust' grounds and that the tests were not invariant. Based on their analyses of the PISA 2006 reading data consisting of 20 items to which all students have responded, and using a different model which takes differential item functioning and local dependence into account, some countries would occupy a broader range of ranks. For instance, Canada would rank between 3rd and 29th if DIF and local dependence were considered instead of 6th and 13th when the Rasch model is used; France would rank

somewhere between 2nd and 42nd instead of 10th and 23rd. This suggests that PISA rankings are not reliable and depend on the items used.

I do raise questions about the validity of the rankings in PISA; however, I do that on conceptual grounds. My main concern is the extent to which the language versions are comparable when language is part of the construct as argued in Chapter 2 and when bias across countries is unavoidable at a certain level as discussed in Chapter 3. Kreiner and Christensen's (2014) methodology led to a massive reduction of the initial scale going from 103 items to 20 items only. Deleting items will change the entire scale and hence, it is not surprising that they end up with different country rankings.

5.7 Countries selected

This study analyses data of the PISA 2006 cycle of three countries – UK, France and Jordan. The selection of countries was based on the language of administration of PISA and the language skills of the trilingual researcher, whose native language is Arabic and is fluent in French and English. Countries which administered the test in one of those languages were considered. A total of eight countries administered PISA in English (Australia, Canada, Qatar, Ireland, Macao-China, New Zealand, United Kingdom, United States of America), five countries administered the test in French (Belgium, Canada, France, Luxembourg, Switzerland) and four countries administered it in Arabic (Israel, Jordan, Qatar and Tunisia). UK was chosen for English, France for French and Jordan for Arabic. The choice of the countries was based on the familiarity of the researcher with the educational systems in these countries. Note that results of Scotland are reported separately in PISA hence the UK sample analysed in this thesis does not include the Scottish sample. For the sake of completeness of

information, it is worth adding that the UK also administered a Welsh version of PISA 2006 tests and questionnaires in addition to the English version. UK and France are both OECD countries while Jordan is an OECD-partner country.

The literature identified language proximity and gross domestic product (GDP) as two variables influencing the comparability between language versions of the same test (Grisay & Monseur, 2007). The closer the languages are to the source languages and the higher the GDP, the less differential item functioning there is between the national version and the source versions. English and French are both Indo-European languages, sharing Germanic and Latin influences at different levels. French is a Latin language with Germanic influences while English is a Germanic language with a Latin influence. These two languages share the same alphabet and very often have similar words. Besides, they are the languages of the source versions. Arabic is a Central Semitic language with a distinct alphabet. The countries vary also in terms of economic power. The UK and France are both European countries, members of the OECD and the European Union (EU). They are both developed countries with similar GDP of \$2.5 trillion¹⁵ and \$2.7 trillion¹⁶ respectively in 2013. Jordan is however a middle-eastern developing country with a GDP of \$33.7 billion¹⁷ in 2013.

The mean scores and country rankings on the science scale of all participating countries published by OECD for PISA 2006 are found in the ‘Analysis’ volume (OECD, 2007b, pp. 56-58). Below, Table 5.5 summarises the mean scores, standard errors and ranking ranges of the countries of interest.

¹⁵ 2,734,949,064,749; <http://data.worldbank.org/indicator/NY.GDP.MKTP.CD>.

¹⁶ 2,521,380,958,035; <http://data.worldbank.org/indicator/NY.GDP.MKTP.CD>.

¹⁷ 33,678,459,200; <http://data.worldbank.org/indicator/NY.GDP.MKTP.CD>.

Table 5.5: Mean scores, standard errors and rankings of UK, France and Jordan on the science scale in PISA 2006.

	OECD Mean	SE	Country Ranking
UK	515	2.3	12-18
France	495	3.4	22-29
Jordan	422	2.8	43-47

The OECD average is set at 500 points with a standard deviation of 100 points. The UK mean score on the science scale is 515 (SE = 2.3) and is significantly above the OECD average. France has a mean score of 495 (SE = 3.4) in science yet is not significantly different from the OECD average. Jordan has a mean score of 422 (SE = 2.8) and is significantly below the OECD average. The three countries considered are hence all considerably different from each other with UK ranking 12-18th, France ranking 22-29th and Jordan ranking 43-47th in science in 2006.

CONCLUSION

OECD invests great effort to ensure that PISA assessments are of high quality in terms of the design, format and content of the items as well as the sampling of the populations and the analytical models. Despite all these efforts, the methodology and techniques used for the development of the tests and the analysis of results were fiercely criticised (Goldstein, 2004; Goldstein et al., 2007; Grisay & Monseur, 2007; Kreiner & Christensen, 2014; Prais, 2003;).

For the 2006 science tests, PISA developed a scientific literacy framework based on the most recent research in science education. According to OECD, the use of the scientific literacy framework ensured that students were not asked curriculum-based questions and are rather assessed on their ability to apply and use their knowledge *of* and *about* science in different

situations. This decision was motivated by the need to ensure that students are not advantaged or disadvantaged by the curriculum they followed.

It is however questionable whether in reality science items are not curriculum-based. Where else would students learn what genetically modified crops are, or how to read graphs or control experiments if it is not something they had covered in school? In Chapter 3 (Section 3.6.1), I discussed Nardi's (2008) paper regarding how different curricula constituted a source of bias unless they were commensurate with the PISA framework. This bias threatens the comparability across countries.

The effort of limiting bias was also apparent in the variety of item formats adopted and the translation and adaptation techniques used. Various types of items were used to limit bias arising from familiarity with item type. In addition, two source versions – English and French - were developed to limit cultural bias (see decentering in Chapter 3, Section 3.5.2.c). Countries were also encouraged to contribute items in their national languages to make the test less culturally-centred. The translation and adaptation processes followed a tedious process and comprised several rounds of verification and refinement of items. However, only the English version of the test underwent field trialling, in Australia, where ACER is based. Besides, despite the thoroughness of the translation and adaptation process, empirical evidence pointed to bias across language versions of PISA tests (Grisay & Monseur, 2007; Grisay et al., 2007; Huang et al., 2014; Kreiner & Christensen, 2014; Le, 2009; Oliveri & Ercikan, 2011; Xie & Wilson, 2008).

Perhaps some of the fiercest criticism against PISA centred on the use of the Rasch model (Goldstein, 2004; Kreiner & Christensen, 2014) and the lack of complete transparency in the

pragmatic decisions taken such as deleting certain items (Kreiner & Christensen, 2014). While the Rasch model is an attractive model because of its simplicity and its ability to map people on the same scale based on their total score, there are various conceptual and practical problems associated with using the Rasch model in international comparisons. I covered the conceptual controversies in Chapter 4 and I will discuss practical issues in Chapter 6. In a nutshell, questions around the extent to which data fit the model were brought up. In the next chapter, I analyse data of PISA science 2006 from three countries – UK, France and Jordan and examine the magnitude of data fit to the model. I investigate the extent to which Rasch requirements are violated, particularly that of unidimensionality

CHAPTER 6

Rasch Analysis and Fit of Data to the Model

*‘Es ist zum Mäuse melken’
‘It is to milk mice’*

(Literal translation from German)

INTRODUCTION

The main aim of this research is to investigate whether language has an impact on question difficulty and demands when comparing language versions of the same test. PISA data from three different countries (UK, France and Jordan) administering the test in different languages are used to investigate this question and to illustrate the issues encountered. In Chapter 4, the case for using the Rasch model was made. From the researcher’s pragmatic stance, the Rasch model was adopted for various reasons. On the one hand, a generalised form of the Rasch model has been employed to analyse PISA results ever since the conception of this test until 2012. On the other hand, adopting the same class of models in this research makes comparisons with the results published by OECD easier to make and better founded.

Analyses have been carried out highlighting methodological issues when analysing large scale assessment data such as PISA data (Prais, 2003; Goldstein, 2004; Goldstein et al., 2007). The most recent of these was published by Kreiner and Christensen (2014) where evidence of great misfit of PISA reading 2006 to the Rasch model was provided. While the complete dataset of all participating countries could fit the Rasch model well, this may not hold for some subsamples such as the one selected for this research.

It is therefore crucial to investigate the extent to which the data of the subsample chosen for this research fit the Rasch model before any conclusion about invariance can be established¹⁸. It is hence necessary to check whether any requirement of Rasch modelling was violated before carrying out further analyses, namely DIF analyses (see Chapter 7). Of particular relevance to the overarching research question is the requirement of unidimensionality. In this chapter, a series of analyses carried out on PISA science 2006 data was described. These analyses examined the extent to which the subsample data fit the model at test level and at item level. The requirement of unidimensionality was however investigated at test level leaving the discussion at item level for Chapter 7.

The chapter is divided into eight sections. The first section reviews the formulae of the dichotomous and polytomous Rasch model used in the subsequent analyses to estimate person and item parameters given that PISA tests comprise both types of items. The second section describes briefly the data retrieval process as well as the sampling procedures and provides country proportions in the initial and the reduced sample. Data were re-coded to be readable by the software used. A rationale of the recoding is included in this section. Given the immense size of the initial dataset, the sample was reduced as is recommended when dealing with large sample sizes. Section 6.3 discusses the interplay between power and precision of a test of fit and their relationship to the quality of results. It also introduces the person separation index used as an index to gauge the reliability and power of a test. The precision and power of a test are related to its targeting.

¹⁸ I have carried independent Rasch analyses for each country and results suggest an overall good fit of data to the model for each country. These analyses were not represented here to keep the thesis within word limit.

Section 6.4 presents the distribution of items and persons to investigate whether the items, based on their location, target the adequate group of persons. Section 6.5 details the tests of fit carried out. These included three overall fit statistics and two individual item fit statistics. Although the data fit the model in general, some items showed evidence of significant misfit. Thorough analyses were carried out to understand potential sources of the misfit. Random guessing in multiple choice items was diagnosed and results were reported in Section 6.6. Violations to the requirement of local independence were evaluated and the outcome of the analyses was outlined in Section 6.7. The functioning of thresholds in polytomous items was assessed and results were presented in Section 6.8. The chapter ended with a conclusion which gathered all pieces of information from the series of analyses conducted to determine the extent to which the data of the subsample of interest fit the model and the extent to which solid grounds were founded to carry out further DIF analyses.

6.1 The Rasch model

Chapter 4 provided an overview of the Rasch model. This section simply reviews the algebraic formulations of the model for dichotomous and polytomous ordered items given that the test consisted of both types of items:

The Rasch model for more than two ordered categories takes the form of:

$$\Pr\{X_{ni} = x\} = \frac{e^{\sum_{k=1}^x (\beta_n - \tau_{ki})}}{1 + \sum_{x=1}^m e^{\sum_{k=1}^x (\beta_n - \tau_{ki})}} \quad (\text{Andrich, 1978}), \quad (6.1)$$

where $x \in \{0, 1, 2, \dots, m_i\}$ is the integer responsible for person n with ability β_n responding to item i with difficulty δ_i , and $\tau_1, \tau_2, \dots, \tau_{m_i}$ are thresholds between $m_i + 1$ ordered categories where m_i is the maximum score of item i .

The special case of the formula above is for dichotomous responses where there is only one threshold δ_i and $x \in \{0, 1\}$.

$$\Pr\{X_{ni} = x\} = \frac{e^{\beta_n - \delta_i}}{1 + e^{\beta_n - \delta_i}}, \quad (6.2)$$

Data from the PISA science 2006 cognitive tests in the three selected countries – UK, France and Jordan – were retrieved and analysed for their degree of fit to the dichotomous and polytomous Rasch models. The data were analysed using RUMM2030 software (Andrich et al., 2009) which applies a weighted maximum likelihood method to estimate the model. The following section details the data collection process for this study.

6.2 Data

Chapter 5 describes thoroughly the development, design, administration and data analysis procedures adopted in PISA. This section is limited to a brief description of the data analysed in this chapter. The first part presents the retrieval of data from online PISA databases and recoding procedures to suit the software package. The second part describes the initial sample of three countries of interest. The third part outlines the sampling procedure and the reduced sample.

6.2.1 Data retrieval and recoding

PISA data, including student, school, and parent questionnaires as well as codebooks and results for all countries in all three assessed subjects, are available online¹⁹ for all cycles. Data of PISA science 2006 cognitive items in three countries – UK, France and Jordan – were retrieved.

Recoding the data was necessary to have a dataset readable by RUMM2030 (Andrich et al., 2009). The software requires the coding of responses to be sequential starting with 0 or 1. In this study, recoding affected only the codes of missing data. Missing values were distinguished by different codes in PISA datasets depending on the reason for which they were missing in the dataset and they were scored accordingly. All multiple responses to an item were coded as ‘8’ and all embedded missing data (i.e. missing values because the item was believed to have been skipped by the examinee) were coded as ‘9’. The codes of these response categories in RUMM2030 (Andrich et al., 2009) followed the sequence of item response categories. For instance, for a multiple choice with three alternatives coded as 1, 2 and 3, missing data were assigned the following successive codes; that is, code ‘4’ was assigned for ‘multiple responses’ and code ‘5’ for ‘embedded missing data’. ‘Multiple responses’ and ‘embedded missing data’ were considered incorrect in PISA (OECD, 2009b) and therefore scored as 0 in RUMM2030 (Andrich et al., 2009).

All structural missing data (i.e. missing responses intended by the test design) were coded as ‘7’ and unreached items were coded as ‘r’ in PISA datasets (OECD, 2009b). These two categories were considered as missing values in PISA and therefore were not scored as

¹⁹<http://pisa2006.acer.edu.au/downloads.php>.

incorrect in the analyses carried out in this research. Considering unreached responses as incorrect responses influences both item and person estimates. If unreached responses were considered as incorrect, a larger number of students would be scoring zero on any given item. This would lead to an increase in item estimates (items would look harder) and a decrease in person estimates. Given that this research is concerned with item parameters rather than person parameters, unreached responses were considered a missing value to avoid an overestimation of item parameters. Both categories were marked by a star sign ‘*’ in RUMM2030 (Andrich et al., 2009). A detailed list of the codes per type of item is available in Appendix to 6.1.

6.2.2 Initial Sample

The initial sample of data from PISA science 2006 test comprised 24,377 students with 24,131 available for analysis (extreme scores were excluded). This sample was composed of three country subsamples; the sample retrieved from the United Kingdom (UK) consisted of 13,152 students, the sample withdrawn from France consisted of 4,716 students and that selected from Jordan consisted of 6,509 students. Note that results of Scotland are reported separately in PISA hence, the UK sample analysed in this thesis does not include data from Scotland (OECD, 2009b). The technical and data analysis reports (OECD, 2009a; 2009b) do not clarify why France, with the largest 15-year old population amongst the three countries (809,375 15-olds in France compared to 779,076 in UK and 138,026 in Jordan), has a relatively small sample of participants. However, issues related to data collection in France for PISA 2006 have been reported. For instance, France did not administer the school questionnaire in 2006 (OECD, 2009a, p.462).

The science cognitive test comprised 103 items. These consisted of various types of questions: 33 dichotomous items (items 1 to 31, items 36 and 103), 6 polytomous items with 3 response categories (items 32 to 35, 37 and 38) and 64 multiple choice items (items 39 to 102).

The extremely large dataset has the potential to yield very precise measures. However, the drawback is that the power to detect misfit. To have more meaningful results, the sample size was reduced. The sampling procedure and the resulting sample are described below.

6.2.3 Reduced sample

The extent of the data fit to the model is investigated using various significance tests. These will be described later in this chapter. A common feature amongst these tests however is that they are all a function of the sample size. A reduced sample was used in this thesis to investigate the extent to which data fit the Rasch model. PISA uses subsamples for international calibrations (OECD, 2009b; See Chapter 5, Section 5.6.3).

Using the ‘Random Select/Person Factor Levels (Country)’ function in RUMM2030 (Andrich et al., 2009), a stratified random sampling was carried out to reduce the overall sample and produce a balanced sample of data from every country. This limits the amount of bias within the overall sample produced by weighting one country more than another in the analysis.

The sample size was reduced to make sure there is equal representation of countries within the sample. The stratified random sampling procedure was undertaken in a way to keep as much information from persons as possible. Since the sample in France is 4,716, the country sample size was reduced to the nearest 1,000, i.e. 4,000 per country, making 12,000 people in total. Given that 93 scores were detected to be extreme, only 11,907 were available for analysis.

Table 6.1 below displays the proportions of countries in the initial sample and the reduced sample.

Table 6.1: Proportions of countries in the initial sample and the reduced sample

	UK	France	Jordan	Total
Initial Sample	13,152	4,716	6,509	24,377
Reduced Sample	4,000	4,000	4,000	12,000

This sample size was still large for test of fit purposes but it maintained enough information for more thorough analyses as there were enough persons per class interval.

The large sample size in PISA studies permits carrying out analyses with high precision and great power to detect misfit. The relationship between precision, power and quality of results is complex. The following section explains the meaning of precision and of power of an experiment and discusses their relationship to each other and to experimental quality.

6.3 Precision of the model and the power to detect misfit

Persons and items can misfit the Rasch model. The precision of the model and the power by which misfit can be detected are directly related. These factors affect the quality of the test. The precision and power of this research were investigated using the Person Separation Index (PSI) (Andrich, 1982). The section below describes each of these properties and their interplay. It also describes the PSI and presents its value for this study based on which inferences on precision and power are made.

6.3.1 Precision of the model

Precision can be defined as the degree to which repeated measurements of the same variable under the same conditions lead to the same result. Keeping all other variables constant, precision increases with the number of observations recorded; i.e., as a function of the sample size (Howell, 1992). In educational assessment, this corresponds to an increase in the number of persons and/or the number of test items (Pallant & Tennant, 2007). Moreover, precision was shown to increase with the number of response categories (Linacre, 2002) and the alignment between persons and thresholds (Pallant & Tennant, 2007). PISA testing produces a massive dataset with a large number of participants per country and a large number of items per subject. The level of precision is therefore very high. An adequate alignment between persons and thresholds could even amplify it. This alignment was empirically diagnosed using the item-threshold distribution function provided by RUMM2030 (Andrich et al., 2009) in the next section.

6.3.2 Power to detect misfit and reliability of a test – the Person Separation Index (PSI)

The power of a statistical test is the probability of correctly rejecting the null hypothesis. It is related to how well the test can avoid making Type I (claiming there is a significant difference between two means when there is none in reality) and Type II errors (claiming there is no significant difference when the significance exists in reality). In other words, it is the degree to which deviations from a model can be correctly detected. In this study, the power of detecting misfit was estimated using the PSI.

The PSI is a function of the variance of the persons' locations (i.e. a function of N). It measures the proportion of variance of person estimates relative to the total variance. The PSI $r_{\beta\beta}$ can be computed using the following formula (Andrich, 1982):

$$r_{\beta\beta} = \frac{\sigma_{\beta}^2}{\sigma_{\beta}^2 + \sigma_e^2} = \frac{\sigma_{\hat{\beta}}^2}{\sigma_{\hat{\beta}}^2 + \sigma_{\hat{e}}^2} \quad (6.3)$$

where σ_{β}^2 is the variance of the locations of the persons, σ_e^2 is the average error variance of measurement among the persons, and $\sigma_{\hat{\beta}}^2$ and $\sigma_{\hat{e}}^2$ are the estimated values of σ_{β}^2 and σ_e^2 .

The greater the PSI is, the greater the power to detect misfit of individuals. The PSI also provides an indication of the reliability of the test; i.e. the degree to which the test yields consistent results under the same conditions. The greater the PSI, the greater the reliability of the test is. The reliability of a test gives an indication of the degree of consistency of repeated measurement.

In Classical Test Theory (CTT), the reliability is the ratio of the true score variance to the total observed score variance. Very often, repeated measurement in educational settings imposes technical and practical constraints, such as learning from the test, remembering answers to questions, fatigue, etc. (Traub & Rowley, 1991). In these cases, Cronbach's Alpha is computed by considering separate parts of the assessment as repeated measures. Cronbach's Alpha assumes that the error of measurement is normally distributed and that the error is not correlated with the true scores. Moreover, the calculation of Cronbach's Alpha assumes no missing data. Given the nature of the incomplete design in PISA and the large number of missing values, Cronbach's Alpha could not be computed for PISA tests while the PSI could

be despite the presence of missing data. In the absence of missing values and lack of extreme scores, Cronbach's Alpha and PSI have very close values (Andrich, 1982).

The PSI value of the reduced sample was 0.83. This value was considered high (Hagquist & Andrich, 2004) compared to a critical value of 0.7 considered suitable for group comparisons (Tennant & Conaghan, 2007) suggesting that the scale is working well to separate people (Hagquist & Andrich, 2004) and was highly reliable. Indeed, the PSI value was close to the reliability measure of PISA science 2006 reported in the technical report as 0.856 without weighting (OECD, 2009b, p.215).

6.3.3 Relating precision and power

The precision and power of a measurement are directly related. The greater the precision of an experiment, the higher is the probability of detecting deviations to the model; i.e. the higher the power of detecting misfit. Given the high level of precision of measurement in PISA due to the large number of items and persons, a great power to detect misfit is an expected outcome. This is manifested empirically by the relatively high PSI recorded.

It is worth noting that in large datasets like PISA, the power could be so high that the data would not fit any model. Even after reducing the sample, some differences may well remain significant because the reduced sample is still too large. To highlight the size of the difference without it being confounded by the sample size, two techniques commonly used in this field were adopted: (1) adjusting a chi square value for a smaller sample size and (2) adjusting the significance levels using Bonferroni correction. These techniques are described below.

6.3.4 Adjusted chi square value

One of the simplest ways to decrease the power to detect misfit is to reduce the sample size (Welkowitz et al.'s, 1991). Reducing the sample size lessens the measurement precision making it less sensitive to slight deviations from the model. In this research, to identify significant misfit only, the overall chi square value was adjusted to a smaller sample size keeping the sample size ($n=12,000$) intact. This adjustment is only algebraic: the observed and expected chi square values are assumed to be the same in each class interval in the smaller sample while preserving the information provided by the larger sample (i.e. $n=12,000$). This procedure has an advantage of decreasing the power while preserving all the information provided by the bigger sample.

The size of the smaller sample used to create the adjusted chi square value was determined using a rule of thumb which consists of randomly selecting 10 to 20 persons per threshold (Linacre, 1999; 2002). The test consists of 109 thresholds (33 thresholds for the 33 dichotomous items, 12 thresholds for the 6 polytomous items with 2 thresholds each, and 64 thresholds for the 64 multiple choice items). Applying the rule of thumb, the sample size needed was between 1,090 and 2,180. The midpoint, 1,635, was taken as a sample size for the chi square adjustment. This way, tests of fit are carried out with the larger sample information yet with more conservative significance criteria.

6.3.5 Bonferroni correction

Given the large number of comparisons of means to be made in the analyses on PISA data, Bonferroni t test, also referred to as Dunn's test (Dunn, 1961), was used to adjust the

significance level in item tests of fit. This correction is a simple way of reducing the risk of familywise (*FW*) error rate or, Type I error in multiple hypotheses tests, where significance is detected erroneously. The Bonferroni t test runs a regular t test but evaluates the result against a modified value of t (t') that has been chosen so as to limit the *FW* error rate.

Dunn's (1961) major contribution was to provide tables for the modified critical values t' . The Bonferroni adjustment consists of dividing the required significance level (typically 0.05 or 0.01) by the number of tests or, in this case, test items (Bland & Altman, 1995). In this study, Bonferroni adjustment was performed with a base of $p < 0.05$, a common base used in PISA analyses (OECD, 2007a). Dividing the significance level by the total number of items (103) results in $p < 0.000485$. Hence, after Bonferroni adjustment, only p -values < 0.000485 were flagged as significant. This adjustment level was used in all the analyses when Bonferroni correction was needed.

Logically, at near population sample sizes, the power of any test would be so great that even trivial and non-meaningful differences would be identified statistically as not fitting the model. In contrast, sample sizes could be so small that any statistical test would have so little power to detect significant deviations. In this case, even large discrepancies would not be rejected and hence not identified as misfitting. When analysing large datasets, the compromise of reducing the extremely high precision level and adopting more conservative criteria occurs at the benefit of screening out significant results with trivial effect sizes to only expose meaningful misfit of considerable magnitude.

The reliability and power of detecting misfit remained high after reducing the sample to 12,000. Additional techniques (Adjusting chi square value and Bonferroni correction) were

used to warrant meaningful interpretations of tests of fit. A high power of a test implies a high precision. As discussed earlier in this section, a high precision could be explained by the large dataset but also by a good alignment between items and persons; i.e. a good targeting of the test. This feature enhances the fit of the data to the model because this means that the test was not too difficult nor was it too easy for the examinees. The section below investigates the targeting of the test by examining the alignment of person and item thresholds and draws conclusions regarding the general fit of the data to the model.

6.4 Targeting – Alignment of persons and item thresholds

A good alignment of persons and item thresholds can enhance the precision of the model as discussed in an earlier section. Moreover, diagnosing the alignment of person and item thresholds provides more information about the data fit to the model. Given that Rasch modelling relies on the contact between the person and the item, it is necessary to evaluate whether the items targetted the right sample of individuals. Using the ‘Person-Item Threshold’ function in RUMM2030 (Andrich et al., 2009), a figure depicting the distribution of persons and item thresholds is produced. This figure allows for a visual comparison of these distributions to test whether the alignment of person and item threshold was appropriate; in other words, whether the items targetted the sample adequately.

The targeting of the test provides information about the overall difficulty of the test as perceived by the examinees. For a well-targeted test (i.e. not too easy, not too hard), with a mean item location score set to zero, the mean location of persons should also be around zero (Linacre, 1994). This can be visualised on the figure produced using the ‘Person-Item Threshold’ function. For instance, if most of the items targetted the lower ability classes of the

population, the test would be easy while if the items mostly targetted the higher ability classes, the test would be rather hard. Figure 6.1a below shows the alignment of persons and items in the overall sample.

Based on Figure 6.1a, the person mean is around 0 with a standard deviation of 1 suggesting that the assessment is of moderate difficulty. The figure shows that items are well aligned with the target population. This adequate alignment implies a high level of precision which partially explains the high PSI (i.e. the great power of detecting misfit) noted in the previous section.

Figure 6.1a: Person-Item Threshold Location

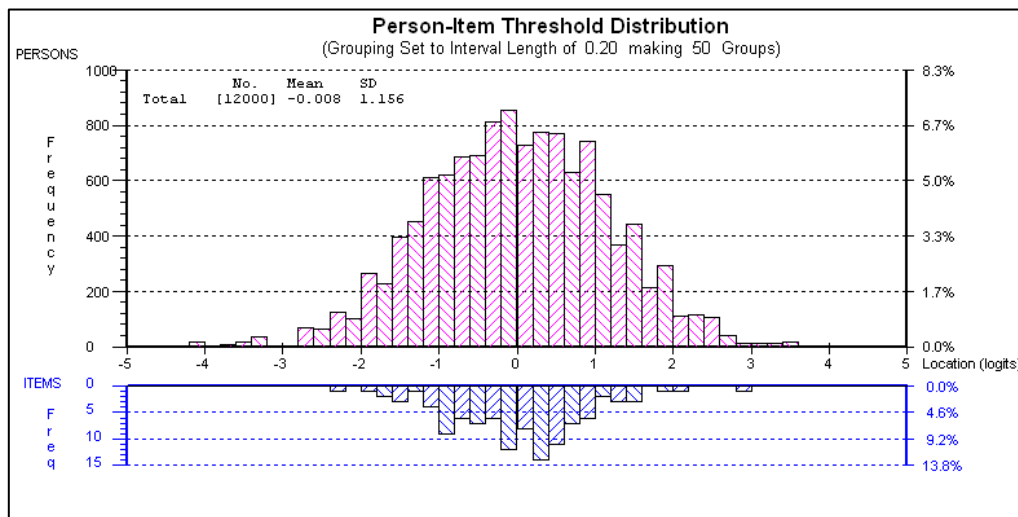
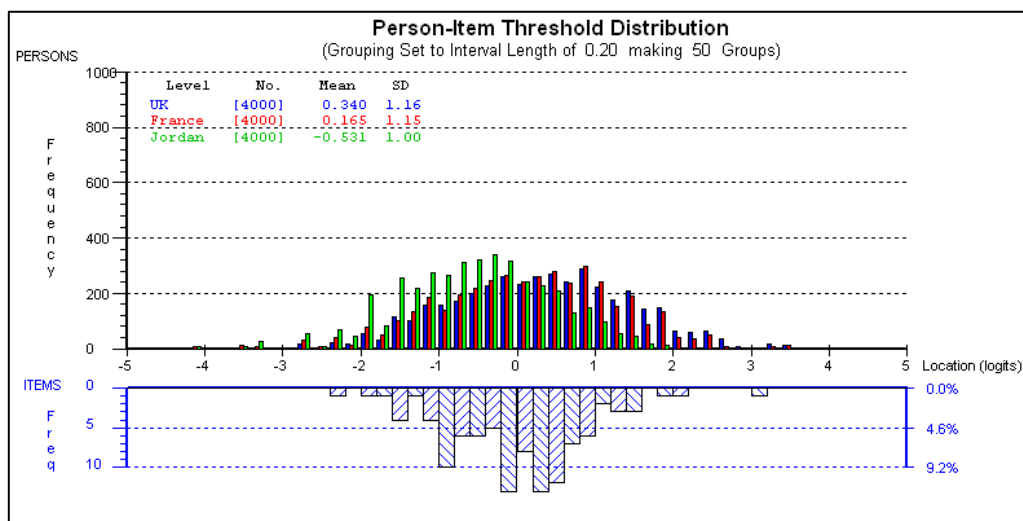


Figure 6.1b below breaks down the general person distribution histogram into distributions of persons per country. Based on Figure 6.1b, the mean of persons in UK is 0.340, in France is 0.165 and in Jordan is -0.531. Hence, students in the UK found the test easier than students in France or in Jordan while individuals in Jordan found the test harder than it was found in the other two countries. The distribution of persons is similar between UK and France (difference in means is 0.175) while it is greater between Jordan and UK (difference in means is 0.871)

and between Jordan and France (difference in means is 0.696). Most of the persons in France and UK are located within the [-1.5; +2] interval while persons in Jordan are located within a lower range [-2.5; +1.5]. Items located above +1.5 logits were too difficult for most participants in Jordan.

Figure 6.1b: Person-Item Threshold Location per country



The results in this section suggest that the test is of moderate difficulty overall. The alignment of persons and item thresholds was adequate. This is an indication that the data fit the model well yet this conclusion needs further bolstering with evidence from tests of fit. The following section summarises the various tests of fit carried out in this study and presents their results.

6.5 Analysis of fit

Various tests of fit were employed to examine the extent to which the data fit the model; some of which were general while others were more specific. General tests of fit included the

alignment of persons to thresholds, item-person interaction statistics and total item-person chi square. Even if the overall fit of the data was satisfactory, particular items could show evidence of misfit. A closer look at individual item fit statistics was carried out by considering fit residuals and chi squares of individual items. The fit residuals were compared to a commonly used cut-off ± 2.5 (Pallant & Tennant, 2007) and interpretations were made accordingly. Chi squares were interpreted using item characteristic curves (ICCs). Results of these tests of fit were used to determine the extent to which the data fit the Rasch model overall and at item level. The following section explains each of these tests of fit.

6.5.1 Overall fit statistics

In this section item-person and item-trait interaction statistics are summarised.

6.5.1.a *Item-person interaction*

Item-person interaction statistics are transformed to approximate a standard z deviate. Hence, the extent to which data fit the model is related to the extent to which the mean is close to 0 and the standard deviation approximates 1.

In this analysis, the mean of the item fit residuals was of -0.1546 with a standard deviation of 5.4067 indicating a good fit of items to the model but a large spread around the mean. The mean of the person fit residuals was of -0.0685 with a standard deviation of 0.8096 indicating a good fit of persons to the model with a low spread around the mean.

6.5.1.b Item-trait interaction

In the Rasch model, the difficulty of items and the ability of persons are mapped on the same continuum. The Rasch model requires symmetry in the ordering of item difficulty and person ability. That is, items perceived as easy by low ability individuals are also perceived easy by higher ability individuals and items perceived as hard by high ability individuals are also perceived as hard by low ability individuals. This translates into a consistent hierarchical ordering of items across the trait reflecting the property of invariance across the trait. The item interaction statistic is reported as a chi square. When a significant item-trait interaction is recorded, this means that the symmetry requirement is violated and that the hierarchical ordering of the items varies across the trait. This could in practice mean that some items perceived as easy by low ability individuals are found hard by high ability individuals.

In this study, the chi square value of the item-trait interaction was of 921 with a non-significant probability of 0.55. This suggests that the requirement of symmetry was not violated and that there is invariance across the trait. In other words, the hierarchical ordering of the items is consistent across the trait.

While results of general tests of fit point to a good fit of the data to the model, a closer look could reveal local misfit at item and person level. Given the overarching research question and the particular interest of comparisons at item level, the next section will be restricted to individual item misfit. General tests of fit indicated a good person fit to the model and this is considered sufficient evidence for the purpose of this analysis. Misfit at item level might be of particular interest to the research question as it may be caused by language features. Only a qualitative analysis of the misfitting items can confirm such speculations yet detecting

individual item misfit is a first step. The following section describes the fit statistics considered and the results of the analysis.

6.5.2 Individual item-fit statistics

Individual item fit statistics are provided as residuals (i.e. sum of individual item deviations) and as chi square statistics. The following formalises each fit statistic, explains ways of interpreting it and presents the results of the analyses.

6.5.2.a Fit residuals

Wright (1980) described the analysis of item and person fit using fit residuals for dichotomous items. Wright and Masters (1982) extended the use of fit residuals to polytomous Rasch models. The following is a brief derivation of the fit residuals for polytomous items.

For an observed response x_{ni} of each person n , of ability β_n for each item i of difficulty δ_i , the residual is $x_{ni} - E(x_{ni})$

$$\text{where the expected value } E(x_{ni}) = \sum_{x=0}^{m_i} x \Pr\{x_{ni}\} \quad (6.4)$$

$$\text{and } \Pr\{x_{ni}\} = 1/\gamma_{ni} e^{-\sum_{k=0}^x \tau_{ki} + x(\beta_n - \delta_i)}, \tau_{0i} \equiv 0 \quad (6.5)$$

and $\gamma_{ni} = \sum_{x=0}^{m_i} e^{-\sum_{k=0}^x \tau_{ki} + x(\beta_n - \delta_i)}$ is a normalising factor ensuring that the sum of all probabilities is equal to 1.

To assess the size of the residual, it is referenced to its standard deviation. Hence, the standardized residual takes the following form:

$$Z_{ni} = \frac{x_{ni} - E(x_{ni})}{\sqrt{V[x_{ni}]}} \quad (6.6)$$

Where $V[x_{ni}]$ is the variance of x_{ni} and is equal to:

$$V[x_{ni}] = E[x_{ni}^2] - (E[x_{ni}])^2 \quad (6.7)$$

The standardised residual z_{ni} approximates a standard normal deviate with a theoretical mean $E(z_{ni}) = 0$ and a variance $V(z_{ni}) = 1$.

However, given that the residuals are formed from the same data where parameters are estimated, the variance of residuals is less than 1, i.e. the degrees of freedom $f_{ni} < 1$.

The sum of the standardised residuals will always be close to 0 because of the estimation equations for the person parameters. Therefore, they are squared to z_{ni}^2 to avoid having a nil sum when assessing their magnitude.

This way, a summary residual y_i^2 for an item i can be compared where residuals over all persons can be added:

$$y_i^2 = \sum_{i=1}^I z_{ni}^2 \quad (6.8)$$

To summarise all item residuals, y_i^2 can be compared to their expected values which are their degrees of freedom $f_{ni} < 1$.

Hence, the item residual is given by:

$$y_i^2 - E[y_i^2] = \sum_{n=1}^N z_{ni}^2 - \sum_{n=1}^N f_{ni} \quad (6.9)$$

Their standardized form can be obtained by dividing them with their standard deviation:

And the item standardized residual is: $Z_i = \frac{y_i^2 - E[y_i^2]}{\sqrt{V[y_i^2]}}$ (6.10)

Given that the standardized residuals are approximate normal deviates, their value should be around 0. The further away the value is from zero, the less the fit of the data to the model. As there is always some degree of misfit, in this analysis, concerns are raised when standardized residuals are below -2.5 or above +2.5. These are standard criteria (Pallant & Tennant, 2007) and correspond to two and a half standard deviations away from the mean, representing a large departure. A positive value for the fit residual is obtained when the data do not conform to the Guttman pattern of responses and that the item discriminates less than expected by the model. A negative fit residual refers to a response pattern conforming well (or even too well) to the Guttman pattern thereby making the item discrimination higher than expected by the model.

6.5.2.b *Chi squares*

Another item level statistic considered in this analysis is presented as a chi square value and was suggested by Wright et al. (1976). The sum of squares of item residuals results in an approximate chi square distribution on $G - 1$ class intervals where G is the number of class intervals. This number can be compared to the values of a chi square distribution on the same degrees of freedom and the associated probability will provide information about how significant is the chi square value; i.e. how likely such a chi square value is to occur by chance. A total chi square estimate for an item could be derived by adding up chi square values of all class intervals of this item.

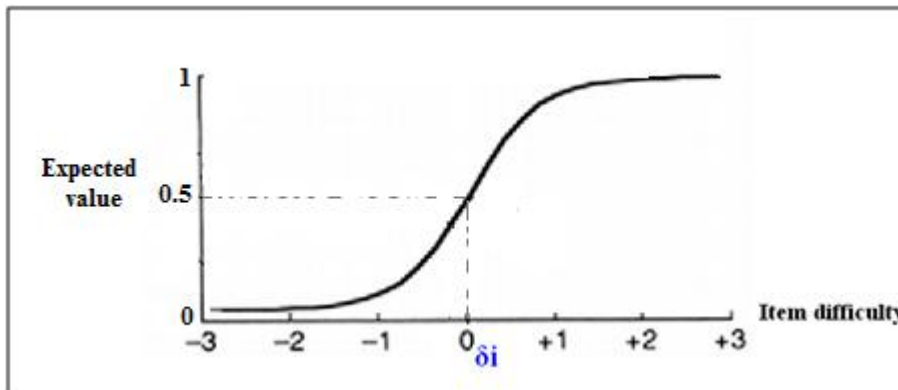
$$\chi^2 = \sum_{n=1}^N Z_{ni}^2 \quad (6.11)$$

Given that this statistic is affected by sample size and the way samples are formed (Smith & Hedges, 1982), interpreting chi square values should be made in conjunction with item characteristic curves (ICC).

6.5.2.c Item characteristic curve (ICC)

The ICC illustrates graphically the theoretical probability that a person v having an ability β_v provides a correct response to an item i of difficulty δ_i (see Figure 6.2 below). The ICC provides the difficulty of the item δ_i by reading the value corresponding to an expected value of 0.5, where the probability of having a correct response is equal to that of an incorrect one. In this case $\delta_i = 0$. The ICC is also the theoretical proportion of the number of correct responses and the theoretical average of the number of persons who got a given item correct.

Figure 6.2: ICC of item i of difficulty δ_i



The ICC allows the comparison between the data (observed proportions) and the model (theoretical proportions represented by the ICC) (Pallant & Tennant, 2007). The further the observed proportions are from the ICC, the more misfitting the data is at item level. When the slope formed by the observed data is steeper than the ICC, the item is considered to be over-discriminating (i.e. the item differentiates between person abilities more than expected by the

model). When the slope formed is flatter than the ICC, the item is considered under-discriminating, i.e. it does not differentiate between person abilities as good as expected by the model).

6.5.3 Results of individual item fit statistics

The results of the individual item fit statistics when running the Rasch analysis on the data are presented below. Conclusions on fit of items were based on information gathered from fit residuals, chi squares and ICCs.

Sixty-two items out of 103 showed fit residuals far from zero (below -2.5 and above +2.5) with eleven items having significant chi square values after Bonferroni correction. Table 6.2 presents these items by increasing order of chi square values.

Based on Table 6.2, items 1, 5, 6, 10, 12, 18, 21, 29, 41, 73 and 96 show strong evidence of misfit with significant chi square values after Bonferroni adjustment. Items 1, 5, 6, 10, 12, 21 and 29 show very negative fit residuals and items 18, 41, 73 and 96 show very positive fit residuals. Large negative fit residuals indicate that items are over-discriminating, i.e. they differentiate between person abilities more than expected by the model. Large positive fit residuals indicate that items are under-discriminating, i.e. they do not differentiate between person abilities as good as expected by the model.

Table 6.2: Item fit based on fit residuals and Chi Square tests after reducing sample to N=12,000 with equal country proportions

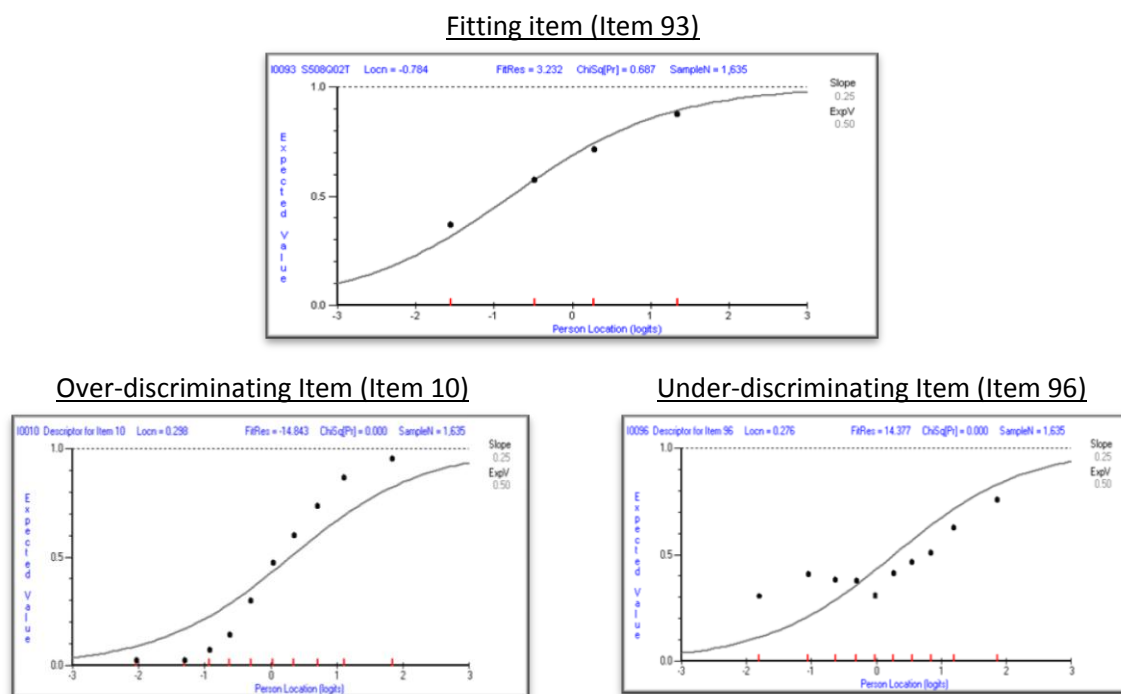
Item	Location	Standard Error	FitResidual	ChiSquare	Probability
93	-0.784	0.038	3.232	1.48	0.686858
86	0.52	0.038	2.872	1.901	0.593188
4	1.206	0.042	-2.521	2.02	0.568353
19	1.352	0.042	-2.713	2.026	0.567093
26	0.913	0.039	-3.101	2.843	0.416541
42	-0.677	0.039	-2.992	2.999	0.391713
33	1.09	0.027	-3.222	3.101	0.376259
102	-0.232	0.038	5.756	3.308	0.346566
72	-0.011	0.037	4.088	3.599	0.308089
89	0.019	0.038	4.069	3.641	0.302981
62	-0.411	0.038	-2.752	3.959	0.26598
24	0.022	0.037	-2.669	4.005	0.260902
94	-1.04	0.04	-3.678	4.141	0.246656
70	0.588	0.039	5.127	4.226	0.238049
17	-0.432	0.039	-4.151	4.541	0.208674
11	-0.262	0.037	-3.143	4.575	0.205690
22	0.451	0.039	-4.928	4.908	0.178655
82	0.313	0.038	6.358	4.919	0.177843
58	-0.897	0.039	3.592	5.257	0.153885
87	-1.782	0.047	5.588	5.677	0.128426
85	-0.864	0.039	-2.985	6.094	0.107104
67	-0.577	0.038	6.451	6.6	0.085803
68	0.492	0.038	3.435	6.818	0.077929
23	1.418	0.043	2.533	7.6	0.055037
81	-1.423	0.044	-5.352	7.91	0.047901
59	-0.86	0.039	-5.101	7.949	0.047073
98	-2.243	0.051	-3.952	8.792	0.032195
64	0.244	0.037	7.796	9.003	0.029251
9	0.447	0.038	-3.924	9.09	0.028115
25	0.681	0.039	-6.704	9.839	0.019988
3	0.189	0.037	-6.504	9.981	0.018726
49	0.125	0.037	8.141	10.022	0.018377
92	0.379	0.038	7.954	10.037	0.018258
20	0.429	0.038	-7.208	10.646	0.013801
32	0.834	0.026	-5.617	10.661	0.013705
56	0.463	0.038	5.815	10.677	0.013605
31	0.738	0.039	-6.037	10.739	0.013225

Item	Location	Standard Error	FitResidual	ChiSquare	Probability
52	-0.447	0.038	9.017	11.238	0.010507
60	-0.791	0.039	6.395	12.444	0.006008
46	-0.094	0.037	-5.834	12.925	0.004803
14	0.787	0.039	-7.557	12.933	0.004785
91	0.419	0.038	8.451	13.048	0.004535
7	0.222	0.038	-8.723	13.122	0.004381
97	-0.068	0.038	9.689	13.266	0.004096
27	-1.431	0.044	-5.591	14.576	0.002218
30	1.443	0.042	3.885	15.495	0.001440
36	0.264	0.038	-9.313	15.934	0.001171
39	0.32	0.038	-8.558	16.057	0.001105
95	0.055	0.037	9.185	16.674	0.000826
101	-0.273	0.038	10.407	16.744	0.000799
75	0.2	0.038	8.66	16.915	0.000737
12	-0.606	0.038	-8.488	17.838	0.000476
41	-1.96	0.047	5.233	18.053	0.000429
6	-0.016	0.037	-10.016	18.19	0.000403
29	0.201	0.038	-9.714	18.824	0.000298
18	0.328	0.038	9.807	19.551	0.000211
21	-0.914	0.04	-6.933	19.795	0.000188
73	0.925	0.04	8.612	21.143	0.000100
5	1.151	0.04	-6.851	22.29	0.000058
1	-0.178	0.038	-11.963	25.669	0.000012
10	0.306	0.038	-15.313	40.128	0.000000
96	0.23	0.037	15.126	41.057	0.000000

Results can be visualised using ICCs. Figure 6.3 illustrates the results using three examples: item 10 (over-discriminating), item 93 (good fit) and item 96 (under-discriminating). The ICC of item 93 suggests a relatively good fit of the observed data to the model. The slightly larger positive fit residual (3.23) could be explained by the slightly flatter slope of the observed data especially for lower class intervals and hence a lower discrimination at this end compared with the model. This difference is however trivial explaining the insignificant chi square value. The ICC of item 10 depicts an over-discrimination: the slope formed by the observed data is steeper than that of the ICC. This explains the very negative fit residual (-15.31). This

severe over-discrimination resulted in a significant chi square value and a significant misfit of the item to the model. The ICC of item 96 portrays the under-discrimination: the slope formed by the observed data especially that provided for lower ability class intervals is flatter than the ICC. This corroborates the very positive fit residual (15.13). The under-discrimination is so strong explaining the significant chi square. Item 96 presents a typical curve for an item suspected to have involved random guessing: lower class intervals scoring better than expected while higher class intervals scoring worse than expected. Random guessing in item 96 will be followed up in a later section (Section 6.7).

Figure 6.3: ICCs of items 10, 93 and 96 (N=12,000)



Analyses on individual item fit revealed negative and positive fit residuals in 62 items indicating that these items were either over-discriminating (negative fit residuals) or under-discriminating. Eleven items out of the 62 also showed significant chi square values after Bonferroni adjustment suggesting significant misfit. Data are considered to adequately fit the

model when misfit does not exceed 15% (Adams & Wright, 1994). In this case, misfit is 11% ($\frac{11 \times 100}{103} = 10.70\%$). Given the size of the dataset, this figure remains acceptable suggesting that the data fit the Rasch model relatively well.

Some items such as item 10 showed misfit because they were highly discriminating. In Section 6.7, high discrimination is reported to be, in some cases, a sign of response dependence (Marais & Andrich, 2008b). Other items such as item 96 were misfitting because they resulted in low discrimination. In Section 6.7, low discrimination is reported to be, in some cases, a sign of multidimensionality (Marais & Andrich, 2008b). The ICC of item 96 also showed potential involvement of guessing which could have contributed to the misfit. The remaining sections of the chapter diagnose items for possible sources of misfit impacting the fit residuals and the chi square values. The next section examines the effect of random guessing in multiple choice items.

6.6 Random guessing

The probability of random guessing is a function of the relative proficiency of a person to the difficulty of the item (Waller, 1976). Persons are more likely to guess the response on an item when the item is too difficult for them. Since there is no random guessing parameter in Rasch, the person proficiency and the item difficulty parameters are susceptible to be affected by random guessing. With random guessing, the item will appear less difficult; i.e. lower item location than expected by the model. Moreover, the number of low ability persons will be higher than expected by the model because they tend to perform better than expected while the number of high ability persons is lower than expected because they are less likely to guess.

These features translate graphically into having the ICC further to the left of the scale than it would be without random guessing; the observed proportions of successes of the low ability persons above the ICC and the proportions of the higher ability persons below the ICC.

Investigating whether random guessing is involved in the PISA test is relevant to the research question in many respects. First, it constitutes an additional diagnosis of data fit to the model as its presence could significantly alter person and item parameters. Second, this investigation helps understand sources of misfit observed at item level in the previous section. Given that guessing from the Rasch modelling perspective, is related to the difficulty of the item, finding that random guessing is not consistent across countries for a particular item could point to two possible reasons amongst others: differences in abilities or differential item functioning. In Chapter 7, thorough DIF analyses and distractor analyses are carried out. These will be more meaningful when combined with evidence on the involvement of random guessing.

Andrich et al. (2012) developed a method to investigate and handle random guessing in a dataset. The following describes this method and its application in this study.

6.6.1 Impact of random guessing on item parameters

Andrich et al. (2012) applied Andersen's theorem (1995; 2002) in which the variance of estimates from a sample of responses is related to the estimates of the same parameters from independent subsets of these responses. Their method consists of two analyses: an initial analysis followed by a tailored analysis. The first one is a typical estimation of person and item parameters of a given dataset where random guessing may have been involved. The

second analysis is a tailored analysis of a subset of responses in which responses believed to have been generated by random guessing in the initial analysis are treated as missing values.

The second analysis is called ‘tailored’ in reference to tailored (adaptive) testing. In tailored testing, items are generated based on a continuous estimation of person ability throughout testing. Responses generated by random guessing are considered to have resulted from a mismatch between person proficiency and item difficulty. This is why in a tailored analysis such responses are treated as non-administered items for these persons.

Using Andersen’s theorem (1995; 2002), the difference in item parameters between the initial set of responses and the subset for each item can be examined. If the ability of a person relative to the difficulty of an item in the initial analysis is such that the probability of a correct response is greater than a cut-off value due to random guessing, then that response is treated as missing. The cut-off value is $1/D$ which corresponds to the probability of randomly guessing an alternative amongst D alternatives by a person who has zero probability of responding correctly to an item. Hence, for a multiple choice item with 4 distractors, the cut-off value would be $1/4 = 0.25$.

Item estimates are generated for the tailored analysis. The difficulty of items affected by random guessing is expected to increase after being tailored. It is worth noting here that every analysis has an identifying constraint: the mean of item difficulty estimates is equal to 0 (*i. e.* $\sum_i^I \hat{\delta}_i = 0$) (Andrich, et al. 2012). Hence, any increase in the difficulty estimates of the items affected by guessing will result in a decrease in the difficulty estimates of the items which have no guessing.

In PISA science 2006 test, items 39 to 102 are multiple choice items. These were analysed for evidence of random guessing using Andrich et al.'s (2012) method. A first analysis estimated person and item parameters from the main sample of responses of all items (including dichotomous and polytomous items). The second analysis – *the tailored analysis* – was performed on a subset of responses – the multiple choice items. The cut-off value was set at 0.3 because the set of multiple choice items involved some questions with 4 distractors (probability = $1/4 = 0.25$) and others with 3 distractors (probability = $1/3 = 0.33$).

To assess the significance of random guessing and the extent to which it has impacted item parameters, a comparison between the difficulty of items before and after tailoring is carried out. A direct comparison cannot be established because the estimates are not generated by the same Rasch analysis. A third analysis – *Average Item Anchoring* – is thus necessary (Andrich et al., 2012). In this analysis, the mean difficulty of a subset of responses to items believed not to have involved random guessing is fixed and the parameters of the remaining items are estimated.

For PISA science 2006 tests, the mean difficulty of dichotomous and polytomous items was fixed because these items were not affected by random guessing. The fixed mean constitutes an additional identifying constraint which, on the one hand, allows the direct comparison of the difficulties before and after tailoring, and on the other hand, limits the extent to which the difficulty of the items not affected by random guessing decreases as a result of the general constraint $\sum_i^I \hat{\delta}_i = 0$. Later, a *t*-test was carried out comparing the locations of the tailored and anchored analyses to investigate which items were significantly impacted by random guessing. Results of the tailored and anchored analyses are displayed in Table 6.3 below.

Table 6.3 suggests that 15 items (shaded in grey) have significantly involved random guessing. The impact is particularly considerable for items 75, 56, 91, 96 and 73. Their location after tailoring showed a change of 0.23, 0.24, 0.25, 0.37 and 0.45 logits respectively. These changes cannot be considered as trivial because that they are changes on a logistic scale. The speculation made in the previous section about item 96 involving random guessing is confirmed based on this analysis. Item 96 but also items 75, 56, 91 and 73 were detected as under-discriminating misfitting items with very positive fit residuals (15.13, 8.66, 5.815, 8.61 and 8.45 respectively) and significant chi squares (Table 6.2).

Table 6.3: Item locations after the tailored analysis and the anchored analyses of multiple choice items in increasing order of t-statistic

Item	Location Tailored	Location Anchored	A – T	SE diff	t = (A-T)/SEdif
39	0.111	0.32	-0.209	0.023812	-8.777175
98	-2.45	-2.243	-0.207	0.040694	-5.0867475
59	-0.936	-0.86	-0.076	0.017664	-4.3026527
85	-0.935	-0.864	-0.071	0.017664	-4.0195835
94	-1.12	-1.04	-0.08	0.021633	-3.6980013
74	-0.949	-0.885	-0.064	0.017664	-3.6232865
81	-1.522	-1.423	-0.099	0.027331	-3.6222206
62	-0.471	-0.411	-0.06	0.017664	-3.3968311
66	-1.594	-1.512	-0.082	0.025612	-3.2015621
46	-0.154	-0.094	-0.06	0.019621	-3.0578831
53	-0.919	-0.872	-0.047	0.019875	-2.3648266
61	-0.13	-0.086	-0.044	0.019875	-2.2138803
77	-1.278	-1.23	-0.048	0.021909	-2.1908902
100	1.97	2.165	-0.195	0.095053	-2.0514953
84	-0.838	-0.807	-0.031	0.015199	-2.0396503
69	-0.156	-0.12	-0.036	0.017664	-2.0380987
80	-1.573	-1.518	-0.055	0.027331	-2.0123448
71	-0.293	-0.262	-0.031	0.017664	-1.7550294
40	-1.167	-1.136	-0.031	0.019875	-1.5597793
99	-0.453	-0.427	-0.026	0.017664	-1.4719601
87	-1.817	-1.782	-0.035	0.028983	-1.2076147
42	-0.698	-0.677	-0.021	0.017664	-1.1888909
45	-0.659	-0.639	-0.02	0.017664	-1.132277
55	0.048	0.067	-0.019	0.017664	-1.0756632
54	-0.914	-0.896	-0.018	0.017664	-1.0190493
76	-0.718	-0.7	-0.018	0.017664	-1.0190493
78	-1.026	-1.009	-0.017	0.019875	-0.8553628
58	-0.903	-0.897	-0.006	0.017664	-0.3396831
79	-0.046	-0.043	-0.003	0.019875	-0.1509464
83	-0.051	-0.056	0.005	0.019875	0.2515773
41	-1.943	-1.96	0.017	0.031749	0.53544967
47	1.445	1.418	0.027	0.048	0.5625
67	-0.565	-0.577	0.012	0.017664	0.67936622
88	0.559	0.538	0.021	0.028983	0.72456884
65	-0.014	-0.03	0.016	0.017664	0.90582163
57	-0.49	-0.507	0.017	0.017664	0.96243548
86	0.548	0.52	0.028	0.025612	1.09321633

Item	Location Tailored	Location Anchored	A – T	SE diff	t = (A-T)/SEdiff
48	-0.06	-0.08	0.02	0.017436	1.14707867
90	0.339	0.305	0.034	0.025612	1.32747698
93	-0.754	-0.784	0.03	0.017436	1.720618
63	-0.889	-0.921	0.032	0.017664	1.81164325
50	0.807	0.738	0.069	0.028983	2.38072618
52	-0.406	-0.447	0.041	0.015199	2.69760195
43	0.035	-0.013	0.048	0.017664	2.71746488
44	1.093	0.98	0.113	0.037947	2.97781146
51	0.773	0.668	0.105	0.032125	3.26850723
60	-0.74	-0.791	0.051	0.015199	3.35555364
102	-0.161	-0.232	0.071	0.017664	4.01958347
82	0.429	0.313	0.116	0.025298	4.58530261
89	0.124	0.019	0.105	0.019875	5.28312336**
70	0.759	0.588	0.171	0.030578	5.59229799**
72	0.104	-0.011	0.115	0.017664	6.51059295**
49	0.241	0.125	0.116	0.017664	6.5672068**
97	0.075	-0.068	0.143	0.019875	7.19511086**
101	-0.143	-0.273	0.13	0.017664	7.35980072**
68	0.695	0.492	0.203	0.027331	7.42738159**
95	0.204	0.055	0.149	0.019621	7.59374315**
92	0.549	0.379	0.17	0.021909	7.7594029**
64	0.4	0.244	0.156	0.019875	7.84921184**
56	0.702	0.463	0.239	0.025612	9.33138227**
75	0.429	0.200	0.229	0.023812	9.61709604**
73	1.376	0.925	0.451	0.039332	11.4665135**
91	0.671	0.419	0.252	0.021909	11.5021737**
96	0.596	0.23	0.366	0.019875	18.4154586**

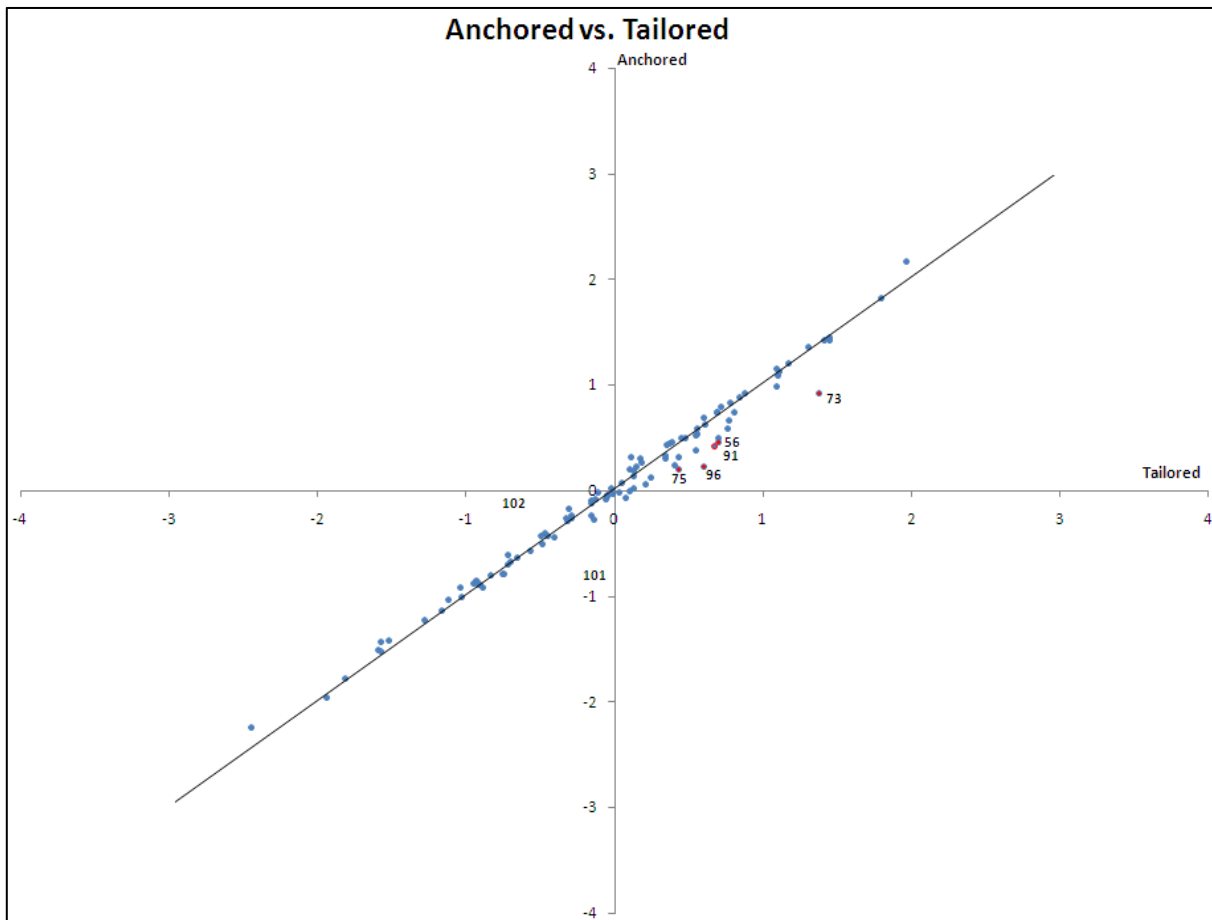
*A-T refers to the difference in item locations between the tailored and anchored analysis

** Significant after Bonferroni correction (base; 0.05), p -values < 0.000485, (critical value: 4.781)

Plotting the mean item locations after the anchored analysis against those after the tailored analysis as shown in Figure 6.4 below reflects the relative impact on item parameters. Five points deviated noticeably from the identity line ($y=x$). These points, marked in red on the graph, corresponded to the five items previously flagged in Table 6.3 for having most significant t values. Random guessing did occur in few items and this is particularly visible in five items marked in red – items 56, 75, 73, 91 and 96 – which deviate the most from the

identity line in Figure 6.4. These correspond to the items with the most significant item location change after tailoring.

Figure 6.4: Anchored versus Tailored mean item locations



Another analysis, *Individual Item Anchoring*, was carried out to investigate whether random guessing significantly impacted person parameters. Details are described in the following section.

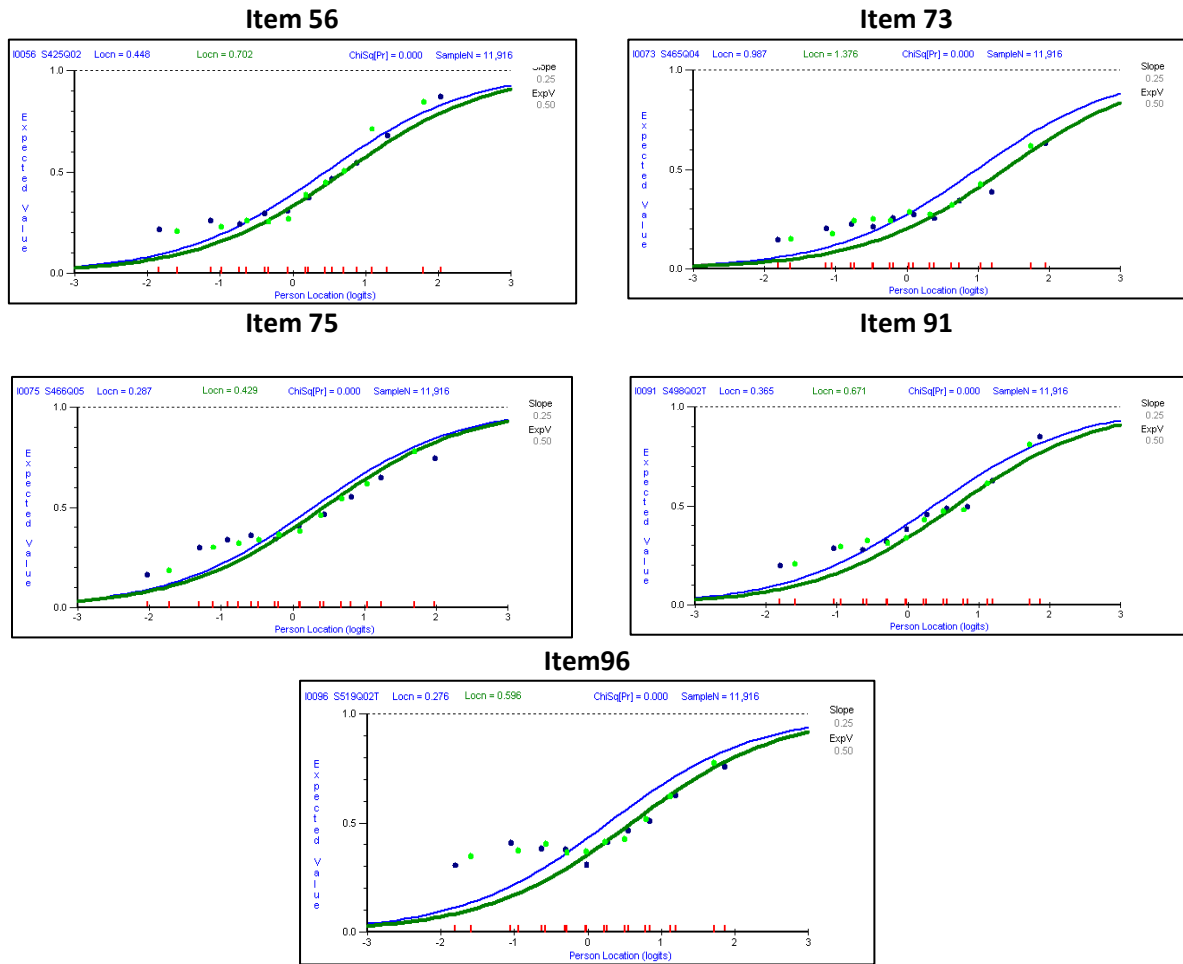
6.6.2 Impact of random guessing on person parameters

The impact of random guessing on the person parameters was investigated by running an *Individual Item Anchoring* analysis allowing the comparison of person parameters before and

after tailoring. From Andrich et al.'s (2012) perspective, random guessing increases the ability of persons who have guessed because they earn credit for item(s) they initially do not know how to answer. In *Individual Item Anchoring* the difficulty of all items in the tailored analysis is fixed while person parameters can vary. The ICCs generated as a result of this analysis are compared with those of the initial analysis.

In this study, only the ICCs of items suspected to have involved random guessing based on the t test evidence (i.e. items 56, 73, 75, 91 and 96) are generated. Results are displayed in Figure 6.5 below. Blue curves designate initial ICCs and green curves represent ICCs generated after the anchored analysis.

Figure 6.5: ICC of items 56, 73, 75, 91 and 96



The initial ICCs (blue ICC) provide additional evidence of random guessing following Andrich et al.'s (2012) visual indices. The observed data of lower ability individuals above the ICCs while those of higher ability students are below the ICCs. The ICCs generated after the anchored analysis (green ICCs) mark a shift to the right, indicating an increase in the item difficulty and a small change of person estimates.

6.6.3 Conclusion

Random guessing could affect person and item parameters and hence the overall fit of the data to the model. It could also explain item misfit detected in items. Investigating the impact of

random guessing on person and item estimates is necessary before carrying out further analyses. In this analysis, random guessing affected the estimates of some items and hence altered person estimates. Despite it being a relatively small change in item and person parameters, the presence of random guessing explains the misfit of individual items (items 56, 73, 75, 91 and 96 detected in Table 6.2. The results of this analysis could be useful when interpreting results of DIF analyses and distractor analyses in Chapter 7.

6.7 Local independence

The Rasch model, be it in its dichotomous form or in its more general polytomous form, both described earlier in this chapter, requires that the data fulfils the requirement of local independence implied by the following formula:

$$\Pr\{((x_{ni}))\} = \prod_n \prod_i \Pr\{x_{ni}\} \quad (6.12)$$

where $((x_{ni}))$ denotes the matrix of responses $X_{ni} = x, n = 1 \dots N, i = 1 \dots I$.

Local independence means that a response on a given item should not govern the response on another item and that the test should measure a unidimensional latent trait.

Instances of response and trait dependence have been described in the literature. Marais and Andrich (2008a) summarised the various examples of response dependence. These included: a correct answer on an item gives a clue to another item, the structure of questions based on a particular logic which provides a hint to an answer (e.g., *Reading one book* and *Reading several books*, and a halo effect operating when raters make decisions on several criteria. Trait dependence occurs when items are associated to the same stimulus (i.e. stimulus dependence,

for example same text, graph, table), cover the same content (algebra, geometry, calculus), or are linked to a common stem (in multiple choice items). Andrich (1985) referred to such groups of items as ‘subtests’. However, other references also exist: testlets (Wang et al., 2002) and item bundles (Wilson & Adams, 1995). Trait dependence is often found in social measurement scales measuring different aspects of a single construct.

The requirement of local independence can be breached in practice by violating any of the two conditions. Marais and Andrich (2008a) named the effects of responses on others as ‘response dependence’ and departures from unidimensionality as ‘trait dependence’. They held that often the generic term of local dependence refers to the former kind (response dependence) while the term multidimensionality is used for the latter type of dependence (trait dependence).

As described in Chapter 5, items in the PISA science tests are organised around units; each unit comprising one or several stimuli (text, diagram, graph, table, etc.). Thus, it is not surprising to find some items correlating because they are linked to the same stimulus. Students struggling with a given stimulus, for example failing to properly read a graph, will not be able to answer items related to it. Often, it will be a case of trait dependence because the stimulus is tapping into an additional trait, for instance ‘reading a graph’, that was measured in addition to the intended trait, for example ‘understanding biological mechanisms’. Moreover, given the design, it is likely to have a response to a question giving a clue to the response to another question.

Violations of the requirement of local independence threaten the fit of the data to the Rasch model. Despite a good overall fit of the data, individual items showed evidence of misfit.

Some items resulted in very negative fit residuals while others yielded very positive ones (See Section 6.5.2). In some instances, large negative fit residuals can be evidence of response dependence while very positive fit residuals can hint to multidimensionality (Marais & Andrich, 2008b). These hypotheses can be tested hypothetically by running further analyses. Moreover, some studies have reported opposite effects of response and trait dependence on the unit of scale and the reliability of the test (Marais & Andrich, 2008a).

This section describes the effect of response dependence and trait dependence on the scale and its reliability as well as on individual item fit statistics. It adopts Andrich and Kreiner's (2010) method to detect and evaluate the magnitude of response dependence and uses principal component analysis (PCA) to investigate potential violation to unidimensionality.

6.7.1 Impact of violations to location independence on the unit of scale and reliability indices

Marais and Andrich (2008b) noted that although response and trait dependence are conceptually disparate, the distinction between them remains blurry in the literature. This is due to the similar effect they both have on data when they are accounted for. A standard procedure to account for the presence of response and trait dependence consists of combining dependent items into higher order polytomous items and comparing the reliability estimates (PSI or Cronbach's Alfa) before and after item combination (Andrich, 1985; Zenisky, Hambleton & Sireci, 2002). Response and trait dependence both result in a decrease of reliability estimates.

However, Marais and Andrich (2008a) found opposite effects of response and trait dependence on the unit of scale and on reliability indices in simulations comparing the effect of each form of violation to independence with respect to a case of absence of dependence. Relative to a case of complete independence, response dependence results in an increased scale (greater range and standard deviation) and an increased reliability (greater PSI/Cronbach's alpha) whereas trait dependence decreased the reliability estimate and reduced the variance in person locations. These features remain harder to detect when data are real because the two forms often compensate for each other in practice.

Another indication is observing item fit residuals. Violation of the requirement of independence introduces some misfit into the model that can be sometimes detected in item fit residuals. Response dependence tends to produce response patterns that are more Guttman-like than predicted by the model and typically result in over-discriminating items; i.e. very negative item fit residuals < -2.5 (Marais & Andrich, 2008b). Trait dependence is manifested in different ways in the data. It acts as 'noise' and creates additional variation thereby making ICCs look less Guttman-like (Marais & Andrich, 2008b), in other words, items were under-discriminating with very positive fit residuals (> 2.5).

In addition, Marais and Andrich (2008b) distinguished each manifestation of location dependence statistically by defining distinct formulae of response and trait dependence. These formulae are presented below.

6.7.2 Formalising local dependence

Marais and Andrich (2008b) formalised algebraically response dependence for dichotomous items as:

$$Pr\{X_{nj} = 1/X_{ni} = 1\} = \frac{e^{(\beta_n - (\delta_j - d))}}{1 + e^{(\beta_n - (\delta_j - d))}}, \quad (6.13)$$

$$Pr\{X_{nj} = 1/X_{ni} = 0\} = \frac{e^{(\beta_n - (\delta_j + d))}}{1 + e^{(\beta_n - (\delta_j + d))}}, \quad (6.14)$$

where $|d|$ is the magnitude of the dependence.

They also formalised trait dependence. The higher order person parameter β measured in the test can be thought of as the combination of β_n , the main common trait of person n intentionally measured in the test, and β'_{ns} is the distinct trait of person n measured by subset s ($s=1, 2, \dots, S$):

$$\beta_{ns} = \beta_n + c_s \beta'_{ns} \quad (6.15)$$

where c_s expresses the magnitude of a variable of subset s compared to the value of the common variable with $c_s > 0$. Variables β'_{ns} of the subsets are considered to be mutually uncorrelated and uncorrelated to the main variable β_n . Marais and Andrich (2008b) proved also that correlations among the subsets are not equal to 1 and that they can be defined by setting equal variances: $V[\beta] = V[\beta_s] = \sigma^2$ for all subsets.

Replacing β by β_n in the equation for ordered dichotomous responses gives

$$\Pr\{X_{ni}^{(s)} = x\} = [\exp(x(\beta_{ns} - \delta_i))]/\{1 + \exp(\beta_{ns} - \delta_i)\}. \quad (6.16)$$

The effect on variance was formalized as follows:

$$B_{ns} = a_s + b_s \beta_{ns} \quad (6.17)$$

where a_s is the mean and b_s is the standard deviation of B_s . As mentioned earlier, despite being conceptually different, distinguishing response and trait dependence remains challenging in the data. Formalising their impact may provide guidance to improve methods in identifying them in real data. In this research, manifestations of local dependence were sought using the readily available techniques to assess their impact on the fit of the PISA data to the model and attempt to understand the misfit detected in some items (Table 6.2). Analyses are described below along with their results. The first series of analyses aims at detecting response dependence while the second series investigates trait dependence.

6.7.3 Analysis of response dependence

This section comprises three subsections. The first outlines items showing possible response dependence, the second quantifies the magnitude of the response dependence and the last provides a conclusion about response dependence in PISA science 2006 tests.

6.7.3.a *Dependent items*

Response dependence was investigated using a residual correlation matrix to detect possible interactions between items. Pairs of items with a residual correlation above 0.2, a cut-off value common in similar studies (Andrich & Kreiner, 2010), were flagged as significant. This meant

that this pair of items had more in common than what all the other items had and that the response to one item (dependent item) was governed by the response of the other item, the independent item.

Item fit residuals were recorded to verify whether items were over-discriminating as observed by Marais and Andrich (2008b). Moreover, Andrich and Kreiner (2010) noted, that often, the dependent item is more difficult than the independent item (i.e. its location is higher than that of the independent item). Hence, item locations were also observed.

Based on the Residual Correlation Matrix (*not provided because of its large size*), interaction (i.e. correlation between residuals > 0.2) is recorded within five pairs of items. A summary of the correlations is provided in Table 6.4 below.

Based on Table 6.4, the response to item 10 seems to be governed by the response to item 9. Similarly, the response to item 12 is governed by that to item 11. The same logic applies to the other 3 pairs. The response to item 53 governed the response to item 58 and 93. The response to item 62 governed that to item 72. Comparing item locations within a pair does not suggest that there is a big difference in item difficulty. This did not conform to data analysed by Andrich and Kreiner (2010).

Table 6.4: Correlated items

Pairs (correlation)	Item number	PISA code	Location	Fit Residual
A (0.201)	9	<i>S304Q03A</i>	0.447	-3.924
	10	<i>S304Q03B</i>	0.306	-15.313
B (0.234)	11	<i>S326Q01</i>	-0.262	-3.143
	12	<i>S326Q02</i>	-0.606	-8.488
C (0.415)	53	<i>S415Q02</i>	-0.872	-2.112
	58	<i>S426Q03</i>	-0.897	3.592
D (0.261)	53	<i>S415Q02</i>	-0.872	-2.112
	93	<i>S508Q02T</i>	-0.784	3.232
E (0.374)	62	<i>S428Q03</i>	-0.411	-2.752
	72	<i>S465Q02</i>	-0.011	4.088

Based on the negative fit residuals, items in Pair A (53 and 94) were over-discriminating. This conforms to the observation made by Marais and Andrich (2008b) in cases of response dependence. However, Pair B shows different features: item 62 was over-discriminating (negative fit residual) while item 72 was under-discriminating to a great extent (very positive fit residual). This case did not entirely conform to Marais and Andrich's (2008b) results and it is thus less clear whether response dependence may have been involved in Pair B.

PISA codes of these items were examined to verify whether they belonged to the same unit, partially explaining the existence of residual correlations. The stems of the PISA codes italicised in Table 6.4 indicated that items 9 and 10 in Pair A belonged to the same unit and therefore were associated to a common stimulus; so were items 11 and 12 in Pair B. This was not the case for items in Pairs C, D and E. Moreover, based on the value of the correlations, the interactions within Pair C and E were the highest.

6.7.3.b Magnitude of dependence

The magnitude of the response dependence was estimated using the Item ‘Split/Item Dependence’ function in RUMM2030 (Andrich et al., 2009). This function is based on the Andrich and Kreiner (2010) method in which the independent item is deleted and the dependent item is split into two items based on two subtests. The first subtest (S_0) is that of people with incorrect response on the independent item and the second subtest (S_1) is that of people with a correct response on the independent item.

Hence, in this study, item 10 was split into 10S0 (subset of people with response = 0 on item 9) and 10S1 (subset of people with response = 1 on item 9) and items 9 and 10 were automatically deleted. The new item parameter can be expressed as:

$$\hat{\delta}_{10S0} = \hat{\delta}_{10} + \hat{d} \text{ and } \hat{\delta}_{10S1} = \hat{\delta}_{10} - \hat{d} \text{ where } \hat{d} \text{ the estimate of the response dependence.}$$

The same procedure was carried out on the rest of the pairs (B – E). Then, new item parameters for the split items were generated.

Table 6.5 below summarises the individual item fit statistics for the split items.

Table 6.5: Individual item fit statistics for split items

Pair	Item	Location	SE	Fit Residuals	df	Chi Square	df	Probability
A	10S0	0.996	0.057	-7.725	2072.76	143.326	3	0.00000
	10S1	-0.546	0.062	-3.219	1421.82	33.872	3	0.00000
B	12S0	0.168	0.057	-5.334	1517.57	47.97	3	0.00000
	12S1	-1.307	0.061	-1.733	1990.55	12.493	3	0.005874
C	58S0	-0.537	0.12	5.468	333.61	54.512	3	0.00000
	58S1	-1.193	0.111	2.225	537.64	16.059	3	0.001103
D	93S0	-1.106	0.123	0.322	333.61	1.441	3	0.695943
	93S1	-0.67	0.101	0.673	537.64	6.08	3	0.10778
E	72S0	0.612	0.112	5.702	438.04	12.737	3	0.005241
	72S1	-0.309	0.107	0.896	437.08	0.577	3	0.901575

Andrich and Kreiner (2010) estimated the magnitude of response dependence d as being the difference between the locations of the items divided by 2,

$$\text{i.e. } \hat{d} = (\hat{\delta}_{S0} - \hat{\delta}_{S1})/2 \quad (6.18)$$

The value of d is usually positive indicating that a correct response on the independent item i decreases the difficulty of the dependent item j , increasing the probability of answering it correctly. The value of d is also positive when an incorrect response on the independent item i increases the difficulty of the dependent item j increasing the probability of an incorrect response on this item. The value of d is rarely negative indicating that a correct response on the independent item i increases the probability of an incorrect response on the dependent item j .

The significance of response dependence between two dichotomous items is defined as:

$$\hat{\sigma}_d = \sqrt{(\hat{\sigma}_{ji0}^2 + \hat{\sigma}_{ji1}^2)/4} \quad (6.19)$$

Where $\hat{\sigma}_{ji0}$ and $\hat{\sigma}_{ji1}$ are the standard errors of difficulty estimates of the split items j_{i0} and j_{i1} (Andrich & Kreiner, 2010). A summary of the magnitude of dependence within each pair and their respective standard errors are provided in Table 6.6 below.

Results in Table 6.6 suggest that dependence is significant at a conservative 1% level for all pairs of items. The highest dependence was detected within pairs A ($d = 0.77$) and B ($d = 0.74$). This value of response dependence is not negligible given that it is on a logistic scale. Based on the PISA codes (see Table 6.6 above), items in pairs A and B correspond to the same unit. Hence, a qualitative analysis of the items and the marking scheme might give an indication on how the response of one item governs the response to the following one. Unfortunately, none of these items was amongst the released ones and access to these items was not granted.

Table 6.6: Magnitude of dependence d between correlated items:

Pair (items)	Dependence $\hat{d}$	Standard error $\hat{\sigma}_d$	z value = $\hat{d}/\hat{\sigma}_d$
A (9 & 10)	0.771	0.042109975	18.3092
B (11 & 12)	0.7375	0.041743263	17.66752
C (53 & 58)	0.328	0.081732796	4.013077
D (53 & 93)	-0.218	0.079577007	-2.73948
E (62 & 72)	0.4605	0.07744837	5.945897

The dependence in items C, D and E is lower yet still significant. It is more difficult to explain given that the dependent item is not in the same unit as the independent item. This is particularly puzzling in the case of PISA tests because of the incomplete design.

Details of the design were outlined in Chapter 5. Items are organised into units and units are organised into thirteen clusters. In 2006, there were seven clusters of science (four clusters of maths and two clusters of reading). The items are administered to students in booklets. There were thirteen booklets; each booklet consisted of four clusters. Given that items within pairs

C, D and E did not belong to the same unit, the resulting dependence suggests that the items could belong to the same cluster.

Table A1.3 in the PISA technical report (OECD, 2009b, pp. 383-386) shows that the pairs of correlated items belong to different clusters. Interestingly, item 53 belongs to cluster S4 and correlated with items 58 and 93 which are both included in cluster S1. With no access to the items or scoring it is difficult to understand the source of response dependence in this case. The value of d is negative for pair D, indicating that a correct response on item 53 increases the probability of responding incorrectly to item 93.

6.7.3.c Conclusion

This analysis showed significant response dependence on item parameters and helped explain the over-discrimination and misfit (very negative fit residuals) observed in items 9, 10, 11, 12, 53 and 63 in section 6.5.2. The misfit of items 58, 93 and 72 was related to positive fit residuals and could not be entirely linked to response dependence. Positive fit residuals were associated to trait dependence (Marais and Andrich, 2008b). The following analysis of trait dependence could help in explaining misfit in items 58, 93 and 72 as well as other items with positive fit residuals.

6.7.4 Analysis of trait dependence: Multidimensionality

This section describes principal component analysis as the method used to detect trait dependence and later presents the results of this analysis. This analysis is directly related to my research question as it investigates whether the PISA science 2006 tests measure more than one dimension. This dimension could be language along many other variables. However,

in Chapter 2 I hypothesised that language is part of the scientific literacy construct in science assessments. In addition, in Chapter 4 (Section 4.4.3) I hypothesised that scientific literacy is a case of a thick unidimensional index variable. Results of the PCA test can furnish evidence for or against these hypotheses and provide insight into the nature of the construct in PISA science assessments.

6.7.4.a Principal component analysis (PCA)

The PCA of standardized residuals (Linacre 1998a; 1998b; Smith, 2002) is used to determine whether items exhibit signs of multidimensionality and it allows the splitting of items into subsets to perform analogous comparisons to the ones described by Andrich (1985). PCA is an analysis similar to a factor analysis which detects whether any subsets of items have more in common than what they share with other subsets. PCA is more sensitive to trait dependence than other indices (Smith, 2002).

Trait dependence in PISA science 2006 data was investigated using the PCA function in RUMM2030 (Andrich et al., 2009). PCA sorts items according to their residual loadings on different components or dimensions. If the test is unidimensional, no particular pattern can be discerned in the residual loadings. Given that the Rasch model extracts the one factor hypothesised to be dominant, the eigenvalues will be very low reflecting the random variation. If the sorting of residuals patterns into more than one subscale, it is evidence of multidimensionality. Eigenvalues of the additional components (PC1, PC2, etc.) will be relatively high reflecting other dimensions.

To determine whether the additional dimension is significant, a *t*-test can be performed (Andrich, 1985) to compare the equivalence of person estimates between pairs of subsets. This

equivalence conforms to the requirement of invariance in Rasch modelling; the comparison of person estimates is independent of the item used to make this comparison. This means that person estimates should remain equal in the two subsets of the same test.

6.7.4.b Results

The violation to the local independence requirement manifested as trait dependence was investigated using PCA of residuals based on the Rasch analysis. Table 6.7 summarises the results of this analysis.

Table 6.7 Summary of the principal component analysis

PC	Eigenvalue	Percent of residual variance	Cumulative Percent*	Standard Error
PC001	3.008	2.92%	2.92%	0.423
PC002	2.728	2.65%	5.57%	0.383
PC003	2.264	2.20%	7.77%	0.318
PC004	2.093	2.03%	9.80%	0.29
PC005	1.898	1.84%	11.64%	0.265
PC006	1.809	1.76%	13.40%	0.253
PC007	1.716	1.67%	15.07%	0.24
PC008	1.628	1.58%	16.65%	0.228
PC009	1.605	1.56%	18.20%	0.225
PC010	1.578	1.53%	19.74%	0.22
...	...	...	...	...
PC103	-1.981	-1.92%	100.00%	0.017

As mentioned earlier, the Rasch model extracts the single factor hypothesised to be dominant. Based on Table 6.7, the eigenvalues of PC1 and PC2 are 3.008 and 2.728 respectively. These eigenvalues are low; they accounted for less than 6% of the total residual variance. This percentage is not high enough to conclude that the requirement of unidimensionality was violated. The misfit associated to the very negative fit residuals and hence the under-discrimination of some items cannot be explained by multidimensionality at test level.

However, multidimensionality could be an attribute of specific items. Item-level multidimensionality could be manifested as DIF (Marais & Andrich, 2008b). The next chapter carries out a DIF analysis to investigate bias and possible multidimensionality at item level.

The fact that the PCA results point to unidimensionality at test level rejects the counterhypothesis that scientific literacy is a multidimensional variable. The argument I made in Chapter 2 about language being part of the construct irrespective of the examiner's intention still stands. It is likely that the scientific literacy variable is, as argued in Chapter 4, an index variable. As mentioned in Chapter 4, the data fit to the Rasch model and the non-violation of the unidimensionality requirement does not provide any indication on whether the variable is a causal or an index variable (Stenner et al., 2008; 2009). The case for one or the other requires a better conceptual understanding of the nature of the variable and more empirical evidence. The conceptual argument has been made in Chapter 2 about the intricacy of the language element in the scientific literacy construct. Chapter 7 will seek more evidence about the nature of the variable in question.

In the next section, details of the analysis investigating the order of thresholds are presented. The order of thresholds was diagnosed for the sake of completeness to the tests of fit (Tennant & Conaghan, 2007). Reversed thresholds are a sign of an inconsistency in the data; i.e. a misfit of items to the model. Items 32 to 38 with the exception of item 36 are polytomous items with 3 response categories. These did not appear as misfitting in Table 6.2, and the analysis is collecting additional evidence to support the empirical finding.

6.8 Order of thresholds

Andrich (1978) derived the rating response model, a class of polytomous Rasch models, by integrating the polytomous threshold formulation available in psychometrics with the polytomous formulation of the Rasch model. Andrich's formulation allowed the interpretation of the total scores in terms of the number of thresholds passed. Thresholds refer to boundaries on the latent variable, and can be assumed to be distances along the same latent continuum. Dichotomous items have one threshold separating two response categories: 0 and 1. Polytomous items with 3 response categories 0, 1 and 2 have two thresholds: the first separating categories 0 and 1 and the second separating categories 1 and 2. Thresholds function normally when their locations are logically ordered on the continuum. For instance, in a polytomous item with three response categories (0, 1 and 2), the location of threshold 1 should be lower than that of threshold 2. That is, it should be easier to overcome threshold 1 and score 1 than it is to overcome threshold 2 and score 2 on that item.

If the thresholds are disordered in an item, this often points to an inconsistency in the data (van Wyke & Andrich, 2006). In this case, analysing the items and the marking schemes qualitatively as well as student responses can help explain the source of the disorder. However, in some circumstances, a way to get around this problem is to collapse adjacent categories by rescoring them the same way and reducing the maximum score of the item accordingly. It should be noted that rescoring is only useful in specific cases where there are too many categories or the categories are poorly defined. Simply collapsing categories to overcome disordered thresholds can lead to further misfit because the probability of a response in any category is a function of all thresholds. This section examines the order of

thresholds and rescores items with reversed thresholds to investigate whether this procedure improves their fit.

6.8.1 Order of thresholds in polytomous items of PISA Science 2006

The order and proper functioning of thresholds were analysed in polytomous items (i.e. items 32 to 35, 37 and 38). The results are presented in the map of Figure 6.6 below.

Figure 6.6: Threshold map of polytomous items 32, 33, 34, 35, 37, 38

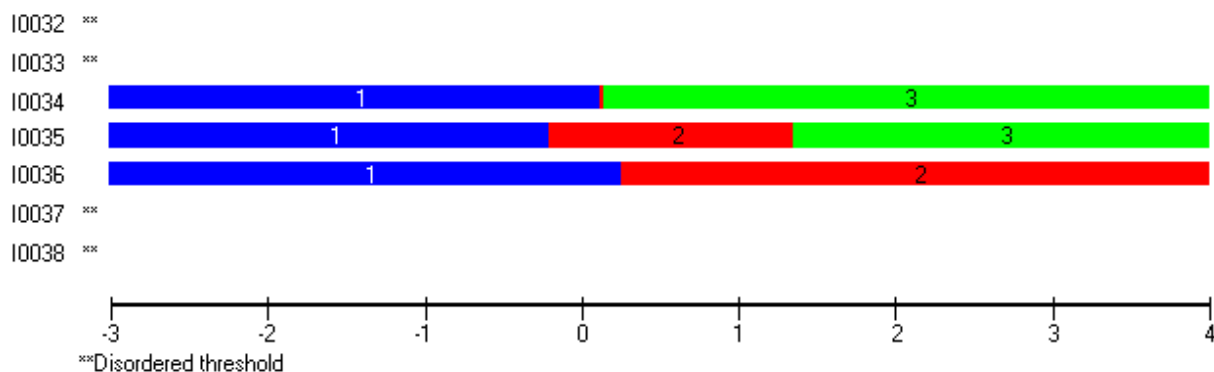


Figure 6.6 suggests that thresholds are reversed in items 32, 33, 37 and 38. Table 6.8 below presents the item location of the thresholds in the polytomous items and provides yet another evidence for the reversal.

The thresholds of items 34 and 35 are logically ordered indicating that the thresholds are likely to be working well in these items. However, items 32, 33, 37 and 38 show reversed thresholds.

Table 6.8: Reversed order of thresholds

Code	Location	Mean	Centralized Threshold 1	Centralized Threshold 2
I0032	0.833691	0.0000	0.143477	-0.143477
I0033	1.090226	0.0000	1.923266	-1.923266
I0034	0.139580	0.0000	-0.012691	0.012691
I0035	0.582369	0.0000	-0.779562	0.779562
I0037	-0.289456	0.0000	0.820904	-0.820904
I0038	0.628915	0.0000	0.339827	-0.339827

The observations made earlier can also be depicted graphically in the graphs of Figures 6.7a and 6.7b showing how categories are functioning.

As predicted, Figure 6.7a suggests that the categories of item 35 seem to be working well. Low ability students have a higher probability of scoring 0, middle ability students have a higher probability of scoring 1 and higher ability students have a higher probability of scoring 2. Item 34 shows that students of ability 0 have roughly an equal chance of scoring 0, 1 and 2. The locations of the thresholds are nearly the same (Threshold 1 = -0.012691 and Threshold 2 = -0.012691 with a difference of 0.0254 logits).

Figure 6.7a: Category probability curves of items 34 and 35

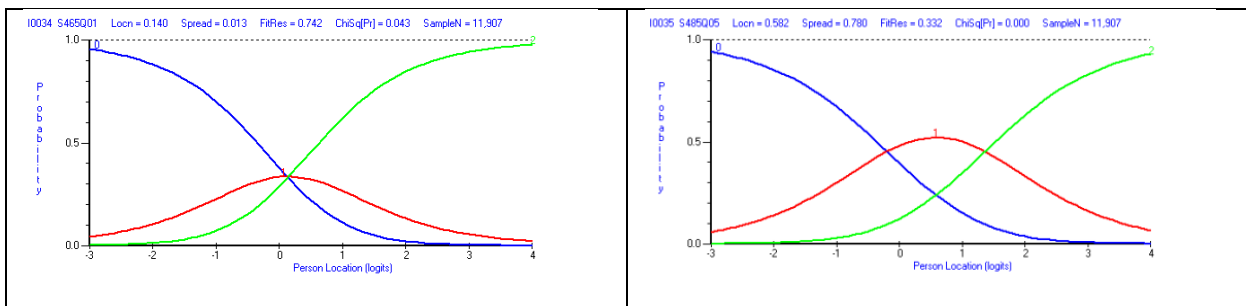
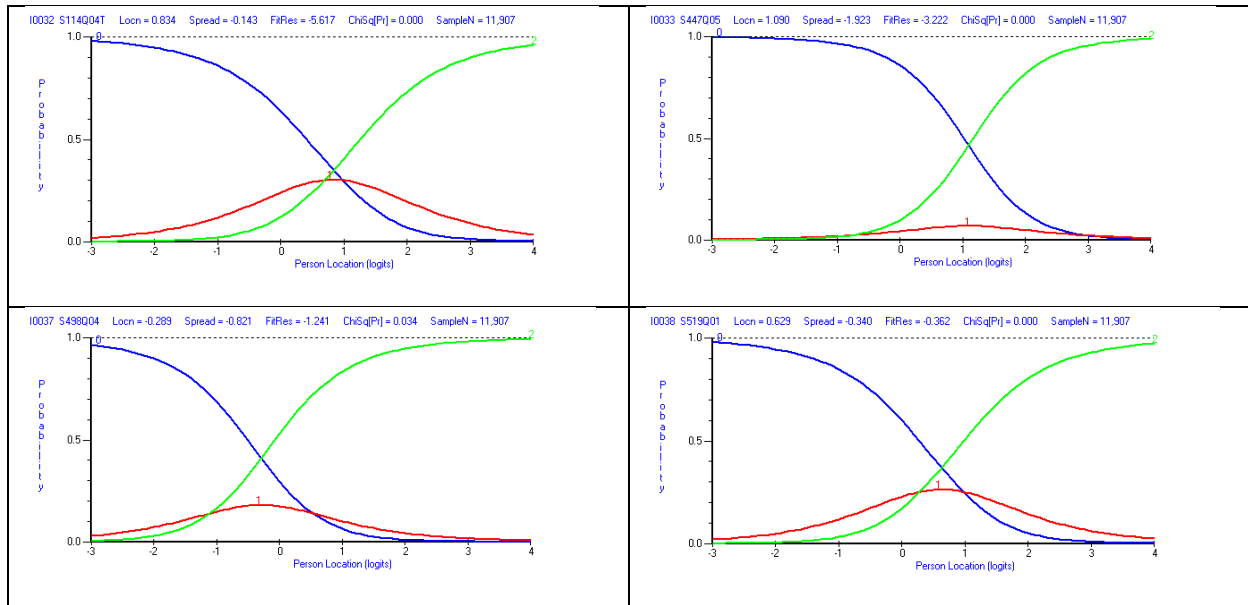


Figure 6.7b (see below) suggests that Category 1 (i.e. corresponding to a score of 1) is not functioning properly in items 32, 33, 37 and 38. At no point on the graph is a score of 1 most

likely for these items and persons with middle ability have a higher chance of scoring 0 or 2 rather than 1. Thresholds 1 and 2 are reversed in all of these items.

Figure 6.7b: Category probability curves of items 32, 33, 37 and 38



One reason for having a category not functioning well is when the discrimination at the thresholds is very low. The discrimination at the thresholds for these items is presented in Figure 6.8 below.

Figure 6.8 provides another visual evidence of the order of thresholds in these items. The thresholds are nearly superimposed in item 34 and logically ordered in item 35 (i.e. Threshold 1 occurs before Threshold 2). Items 32, 33, 37 and 38 show reversed thresholds.

In all items, the discrimination at the thresholds is very similar to the expected curve. Although there is a high discrimination at Threshold 2 in item 33 for low ability students (very steep slope formed by the observed data points compared to the expected value). There is also a low discrimination at Threshold 2 for low ability students in item 38. Hence, the reversal cannot be related to a low discrimination at the thresholds.

Figure 6.8: Threshold probability curves of polytomous items

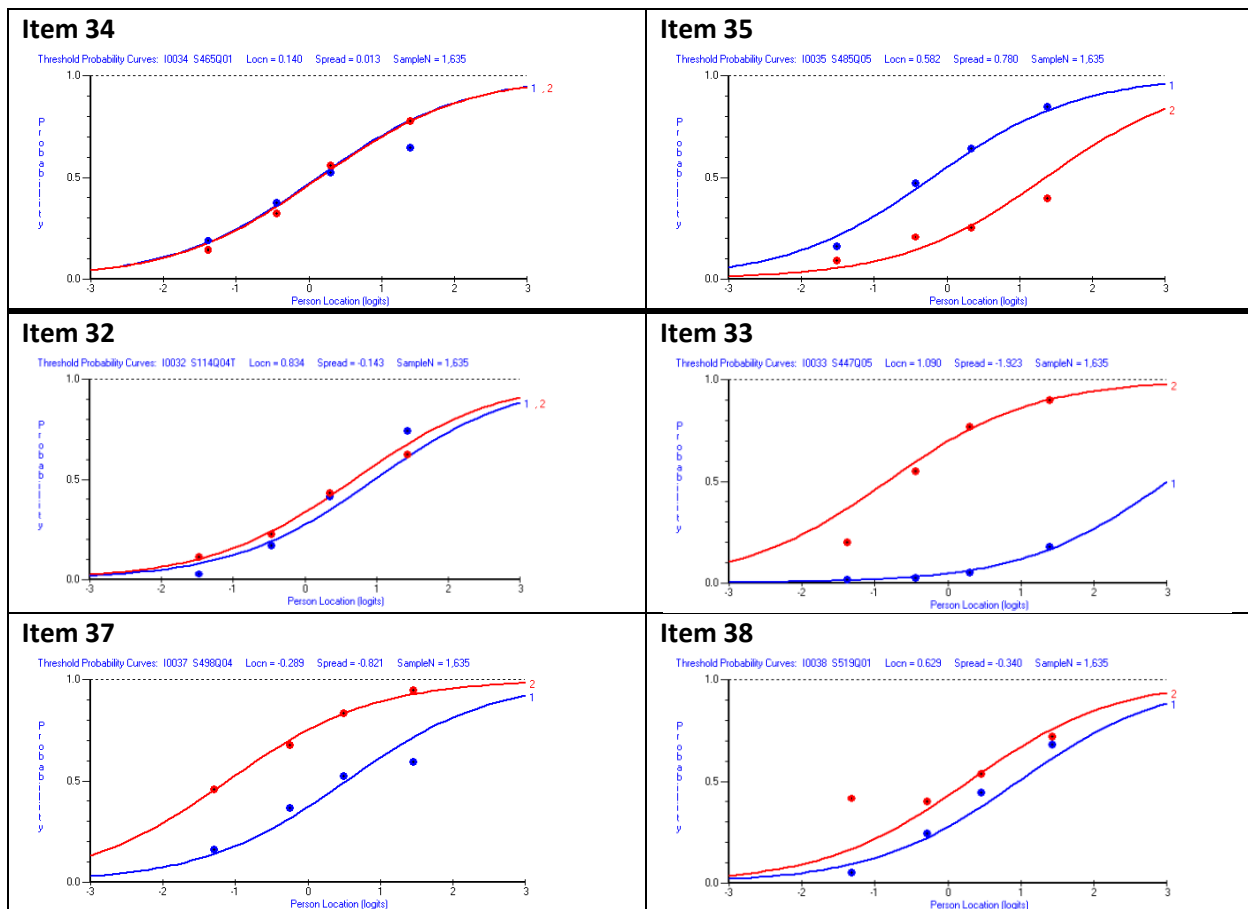


Figure 6.1a in Section 6.4 showed that most persons are located between -2.0 and 2.0 logits. The thresholds of the former items are located within this range (see Table 6.8). This suggests that the reversal in thresholds is not due to the lack of data in this region. It is rather due to a more substantive reason. This makes it reasonable to combine response categories 1 and 2 for all these items and rescore and investigate whether this procedure improves the fit of these items. If rescoring resulted in higher misfit in a particular item, initial estimates with the reversed thresholds would be recovered.

Table 6.9 below shows the fit residuals and the chi square test results of items 32, 33, 37 and 38 before and after rescoring. The results in Table 6.9 suggest that rescoring worsened the fit

of the items with reversed thresholds; their fit residuals moved further away from zero and chi square values became larger with lower probabilities. Initial scores were therefore recovered.

Table 6.9: Fit residuals and Chi Square test results before and after rescoring

Location			Fit Residual		Chi Square Test			
Item	Before	After	Before	After	Before		After	
					χ^2	Prob	χ^2	Prob
32	0.83369	0.393	-5.617	-12.01	10.661	0.013705	30.275	0.000001*
33	1.09023	1.527	-3.222	-7.386	3.101	0.376259	25.029	0.000017*
37	-0.2895	-0.788	-1.241	-7.532	1.192	0.754934	12.799	0.005093
38	0.62892	0.216	-0.362	-8.883	3.534	0.316354	14.269	0.002562

* Significant after Bonferroni adjustment (base $p < 0.05$); Prob: Probability

Increased misfit was expected after rescoring given that discrimination was adequate around the thresholds. Collapsing categories did not solve the problem of reversed thresholds meaning that the source of the malfunctioning of the thresholds did not result from an error in the marking scheme but rather from a more substantive reason worth investigating qualitatively at a later stage. It is however worth noting that the reversal of the thresholds in 4 out of the 6 polytomous items in this test did not translate into any evidence in the individual item fit statistics. Therefore, it is always necessary to conduct all types of diagnoses when investigating the fit of data to the Rasch model. Tests of fit, even when powerful, paint a rough sketch. It is only when delving deeper into the data that more information can be gathered.

CONCLUSION

Rasch modelling requires the data to fit the model so that valid interpretations can be founded. PISA does not publish the fit statistics of the entire dataset in its report (see OECD, 2007a,b, 2009a,b). While data of the entire PISA dataset may fit well the Rasch model, this might not

hold for smaller subsamples such as the one selected for this research. In this chapter, a Rasch analysis was carried out on the data to investigate the extent to which the data fit the model before carrying out any further analyses on the differential functioning of items in the different countries. Data from PISA science 2006 of three countries: UK, France and Jordan were analysed. Using random stratified sampling, the sample across the three countries was reduced to 12,000 in a way to have equal representation of countries.

A high PSI was recorded suggesting a high reliability and power to detect misfit. Results of the tests of fit suggested an overall adequate fit of the persons and items to the model using an Adams and Wright's (1994) 15% criterion as a limit for acceptable misfit. Nevertheless, individual fit statistics showed eleven items out of 103 with strong evidence of misfit. Further diagnostic analyses were carried out to investigate possible sources of misfit in these items. Besides, it was necessary to verify whether the Rasch requirements were violated before carrying out further analyses. The additional diagnostic analyses included examining the impact of random guessing on item and person parameters, the possible violations to the requirement of local independence, and the order and functioning of thresholds.

The results of the analyses investigating the effect of random guessing suggested that random guessing affected five items. These were already flagged for significant misfit and under-discrimination. The analysis on the order of thresholds identified four out of the six polytomous items with reversed thresholds. However, none of these items had shown any evidence of misfit. The rescoring technique did not improve the item fit suggesting that the source of the reversal was not simply associated with problems with the marking scheme but, rather, with more substantive reasons.

Local independence was investigated in terms of response and trait dependence. Results suggest that significant response dependence was detected within five pairs of items. These were flagged for over-discrimination in some misfitting items. Trait dependence was analysed using PCA of residuals. The results suggested that scores could be explained by only one major underlying factor accounted for by the Rasch model; the rest of the variance was random. This constituted evidence that the test is unidimensional. In Chapter 2, I argued that language is an intrinsic part of the scientific literacy construct and in Chapter 4 I hypothesised that scientific literacy is a thick index variable. Evidence for unidimensionality of a test cannot prove the hypothesis right but can provide evidence against considering language as an extricable variable that can be controlled.

In the next chapter, DIF analyses are carried out to investigate whether items behaved differently in different countries. The presence of DIF is a sign of multidimensionality at item level. This evidence can be used to better understand the complexity of the construct of scientific literacy measured by PISA.

CHAPTER 7

DIF and Distractor Analyses

‘The limits of my languages mean the limits of my world.’

(Wittgenstein, 1922)

INTRODUCTION

In Chapter 6, language versions of the same PISA science 2006 test of three countries – UK, France and Jordan – were analysed using the dichotomous and polytomous forms of the Rasch model. Results provided evidence for a good general fit of the data to the model. From the Rasch perspective, this is evidence that the language versions are invariant across the countries. However, additional analyses on individual item fit statistics identified 11 items showing strong evidence of misfit.

In the same chapter (Chapter 6), further diagnoses were carried out in an attempt to explain the detected misfit. Random guessing and violations of the requirement of local independence were investigated. Results suggested that random guessing was involved in few items, partially explaining their misfit. Similarly, response dependence was detected in five pairs of items explaining the over-discrimination and misfit of several items. A PCA analysis indicated that the requirement of trait independence was not violated and that the test could be considered as operating unidimensionally.

Given the emphasis of this research on invariance at item level, additional analyses are required to investigate whether the unidimensionality exhibited at test level would be maintained at item level. Oliveri and Ercikan (2011) found that language versions of the same PISA problem solving test were similar at test level; however, some items functioned

differently in different countries (differential item functioning, differences in item discrimination correlations). This was due to the interaction of the problem solving ability with (an)other, possibly unintended, dimension(s) such as language. If language features advantaged groups randomly (i.e. not consistently in a particular direction), these advantages would tend to cancel each other and balance out in a long test like PISA. Hence, despite having items not functioning the same way in all groups, the effect could be neutralized at test level. However, in some contexts such as test construction from item banks, equivalence at item level is particularly important for its own sake.

Non-equivalence at item level can provide evidence that might be worth following up more substantively for a better understanding of the variable assessed and for improving item construction. This is a challenging question in psychometrics given that bias is a relative measure; it takes into consideration how other examinees have performed, within the same group and in other groups. In other words, statistics cannot guarantee equivalence of items in the absolute. Statistical equivalence of items is always associated to how these items functioned in a particular context across different groups.

This chapter describes the methods used in investigating the extent to which items were invariant across language versions of the same test. DIF analyses were carried out to identify items behaving differently in data from different countries. DIF is often manifested in items where the requirement of unidimensionality is violated (Marais & Andrich, 2008b). It follows that items showing DIF can have at least one additional dimension that may confer advantages to some groups while disadvantaging others. Unfortunately, DIF analyses can provide information about how well an item is functioning relatively to all the other items on a given scale. They do not provide evidence identifying particular factor(s) as causing the bias.

However, in the context of this research, I hypothesise that causal factors of DIF include language difference and translation and adaptation impact on item difficulty. The same item formulated in different languages might be easier or harder to comprehend and answer in some countries.

Previous studies compared language versions of the same international test (PISA, TIMSS, PIRLS) and used common methods such as IRT (Ercikan & Koh, 2005; Le, 2009; Kreiner & Christensen, 2014; Huang et al., 2014) and logistic model regression techniques (Wu & Ercikan, 2006; Hauger & Sireci 2008; Xie & Wilson, 2008; Oliveri & Ercikan, 2011) to detect DIF. In this research, a two-way analysis of variance (ANOVA) of response residuals was adopted to flag items with DIF (Hagquist & Andrich, 2004). This approach has not been used in international assessment yet but was employed in a few studies (Andrich & Hagquist, 2012, 2014; Tennant & Pallant, 2007). In addition to examining bias at item level, DIF was investigated at test level using a new approach that is later described in the chapter.

Some DIF items are multiple choice items. Countries may vary in the frequency with which certain distractors are selected. Some distractors may be more or less attractive to select in a given country. Various factors such as common misconceptions (Herrman-Abell & DeBoer, 2011) could explain that choice and language could be one of them. A distractor analysis was carried out on multiple choice items irrespective of whether they had been flagged for DIF or not. However, as will be explained later in this chapter, only a few items were eligible for this type of analysis. A comparison of the distractor choices in multiple choice items eligible for the distractor analysis across the three countries gave additional insights into possible explanations of DIF and uncovered differences that were not picked up by the DIF analysis.

This chapter is divided into four main sections. The first provides a literature review of similar studies and explores the various approaches used to analyse DIF. The second section describes the methods adopted to detect DIF in this research. The third section summarises the results of the DIF analysis at item and at test level. The fourth section provides the results of a distractor analysis carried out on a subset of items. The chapter ends with a discussion of the analyses carried out in this chapter and links the findings and limitations so far in the thesis to the main overarching argument.

7.1 Differential Item Functioning

Invariance in measurement implies that persons, having the same value of the latent variable (W) measured but from different groups (V) have the same expected value of a response on an item (Andrich & Hagquist, 2012). In other words and in the context of this research, the probability of receiving the observed score Y for a person given her science ability w on the latent trait W and given her country membership v , is equal to the probability of having the same observed score Y for any other person with the same science ability irrespective of the country membership. Millsap & Everson (1993) formalised measurement invariance in the following way:

$$P(Y | W = w, V = v) = P(Y | W = w), \quad (7.1)$$

When invariance is violated, Equation 7.1 no longer stands. The probability of receiving a particular score for two persons with the same location on the latent trait varies depending on the group membership. This is referred to as *differential item functioning* (DIF).

Historically, DIF studies were developed with the increased awareness of providing equal opportunities to ethnic minorities and genders (Angoff, 1993). Detecting bias in measurement was linked to fairness and equity issues. The term *differential item functioning* was introduced to avoid the political and emotional connotation embedded in the term '*bias*' (Holland & Wainer, 1993). Sireci et al. (2005) distinguished three terms: *item impact*, *DIF*, and *item bias*. DIF analyses investigate whether observed statistics indicating item impact are due to 'true' differences in proficiency on a measured construct across groups or whether these differences are due to construct-irrelevant variables introducing bias dis/advantaging a group. Analyses of DIF and item impact are purely statistical while studies of item bias are qualitative in nature. In this thesis, I mainly use the acronym DIF to refer to measurement inaccuracy across groups and incidences of the terms '*bias*' and '*dis/advantage*' are not meant to bare any political nuances.

In this research, I investigate whether invariance is violated at item level and at test level. When DIF is manifested in an item, it implies that the item measures an additional aspect beside the targeted construct for which examinees are not equal. In this research, examinees initially differ in their country membership. Country membership is a complex variable as it can be composed of a very heterogeneous population in terms of culture, curricula, teaching, etc. These aspects could be contributing in different ways and to different extents to the manifestation of DIF. However, for the sake of these analyses, the tests are considered to be the same in all respects except the language of administration to which country membership is used as proxy.

One way of thinking about DIF would be that language is playing the role of a second latent variable W_2 (W_1 being the ability in science) on which persons have different values w_{2a} , w_{2b}

or w_{2c} . Equation 7.1 holds for $W = (W_1, W_2)$ but not for $W = W_1$ (Millsap & Everson, 1993). In more explicit words, the probability of receiving the observed score Y for a person given his/her science ability w_1 and his/her language proficiency w_2 on the latent trait W and given his/her country membership v , is equal to the probability of having the same observed score Y for any other person with the same science ability and the same language proficiency irrespective of the country membership.

This equation might however still be problematic in this particular context because language groups are defined by country membership. Imagine a scenario where two groups of students from the UK and Jordan, both highly proficient in the language of instruction (English and Arabic respectively) are answering the same item in the relevant language. In this case, language proficiency would not be a predictor. Potential bias would be introduced by the way the item was translated, for instance providing an unintended hint to the answer for one language group defined by country membership. Differential item functioning should be interpreted in light of this insight. It is also important to note that DIF is a relative measure and can only be manifested when few items show DIF. DIF cannot be detected if the majority of items are biased in the same direction.

A classical approach to detect DIF has been the Mantel-Haenszel procedure (Mantel & Haenszel, 1959) modified by Holland and Thayer (1988). However, this method is not employed in analysing DIF in international large-scale assessment because it is only appropriate for dichotomous data and cannot handle large amount of missing data. Studies comparing language versions of the same international assessment at item level made use of loglinear item response models (Mellenbergh, 1982; Kelderman, 1989), logistic regression models (Swaminathan & Rogers, 1990) and likelihood-ratio tests (Thissen et al., 1988).

Ercikan and Koh (2005) used 2- and 3-PL IRT models to examine the comparability of constructs between the English and French versions of the science and mathematics tests in TIMSS 1995 in Canada, England, France and USA. Others employed logistic regression models to detect differential item functioning in TIMSS 1999 data in USA and Taiwan (Wu & Ercikan, 2006) and in Singapore, USA and Iran (Hauger & Sireci, 2008). Grisay and Monseur (2007) and Grisay, et al. (2007) used factor analysis along other analyses to evaluate DIF. More recently, Kreiner and Christensen (2014) carried out a DIF analysis based on Andersen's (1973) Conditional Likelihood Ratio (CLR) method on PISA reading 2006 test on all datasets.

A study carried out by Xie and Wilson (2008) employed sophisticated linear logistic test models to detect measurement bias at the level of item facet (an item facet being a common feature shared by a group of items) in PISA maths 2003 tests in two pairs of countries matched on their PISA performance in maths: Canada and Japan; USA and Spain. Huang et al. (2014) used multidimensional random coefficient multinomial logit model (MCRMLM) to investigate DIF in the English and Chinese versions of PISA science 2006. They compared data from four countries (USA, Canada, Hong Kong and China).

The following table, Table 7.1, summarises the main studies, published in English, comparing language versions of the same international test as well as the methodology adopted and the main findings. For methods detecting DIF across language versions of an assessment but not necessarily international tests, refer to Sireci, et al. (2005).

Table 7.1: Studies comparing language versions of the same international test

Study	Test (Year)	Countries (language)	Subject	Aim	Methodology	Main findings
Factor Analysis						
Grisay & Monseur (2007)	PISA (2000)	47 countries (various languages: English, French, German, Spanish, Chinese, Japanese, etc.)	Reading	Explore equivalence issues across the 47 test adaptations in various languages Investigate the relationship between percentage of commonality and country GDP	Between-countries variance in item difficulty and within-countries variance attributable to DIF Factor analysis	Item difficulty was associated to the common factor measured by PISA tests; however, the variance in item difficulty was not random. The amount of DIF increases with the distance between the languages included in the comparisons. The amount of DIF increases for low-GDP countries, also characterised by low average achievement in reading.
Grisay et al. (2007)	PISA (2005 – Field trial), PISA (2006)	All national versions in languages used for $\geq 5\%$ of the target population (77 versions, 42 languages)	Maths, Science	Develop a set of indicators to be used as criteria for evaluating equivalence across countries	Cluster analysis Factor analysis Stepwise regression	Identifying geographic and linguistic patterns of differences Assessing the magnitude of possible linguistic bias Understanding some of the potential sources of bias Impact on item difficulty
Simultaneous item bias test (SIBTEST)						
Sandilands et al. (2013)	PIRLS (2001)	Colombia (Spanish), USA (English), England (English)	Reading	Investigate sources of DIF in international large-scale assessments	Confirmatory approach; Analyse items for differential bundle functioning (DBF) using polytomous simultaneous item bias test (Poly-SIBTEST)	Differences between languages. However no significant DBF detected.

Study	Test (Year)	Countries (language)	Subject	Aim	Methodology	Main findings
<i>Logistic regression</i>						
Wu & Ercikan (2006)	TIMSS (1999) 8 th grade	USA (English) and Taiwan (Chinese)	Maths, Science	Identify cultural sources of DIF	Multiple-variable matching method using logistic regression analysis.	Extra lesson hours after schools are a potential source of DIF
Hauger & Sireci (2008)	TIMSS (1999) 8 th grade	Iran (Farsi), Singapore (English), USA (English)	Science	Detect DIF across examinees tested in their dominant language and examinees tested in a second language	Logistic regression DIF detection, effect size criteria	No significant differences between students tested in their native language and those tested in a non-native language of instruction.
Xie & Wilson (2008)	PISA (2003)	Canada (not specified whether it is English only, French only, or both), Japan (Japanese), USA (English), Spain (Spanish)	Maths	Investigate DIF in international tests using differential facet functioning (DFF) and multidimensional differential facet functioning (MDFF)	Linear Logistic Test Model (LLTM) – DFF & MDFF	DFF model reveals more substantive explanations of DIF MDFF allows individual differences to be modelled on the dimension exhibiting the DFF effect
Oliveri & Ercikan (2011)	PISA (2003)	Canada (English and French)	Problem solving measure (PSM)	Examine construct comparability in multilanguage assessment using different approaches	DIF using latent class modelling (logistic regression analyses)	At test level, the PISA language versions are highly similar. At item level, some items show DIF, differences in item discrimination correlations, linguistic differences.

Study	Test (Year)	Countries (language)	Subject	Aim	Methodology	Main findings
<i>Item response theory models</i>						
Ercikan & Koh (2005)	TIMSS (1995) 8 th grade	Canada (English and French), England (English), France (French), USA (English)	Maths, Science	Investigate comparability of constructs across language	Comparing item parameters using 2- and 3-PL IRT. Comparing item response functions (IRFs) by creating DIF measures.	Differences in constructs across language versions especially across France and USA Limitation to comparison of constructs across countries.
Le (2009)	PISA (2006)	50 countries, (60 language versions)	Science	Investigate gender DIF across countries and test languages	DIF is computed as the difference between relative difficulty estimates (2- and 3-PL IRT model)	Gender DIF for science is dependent on item format and content domains
Kreiner & Christensen (2014)	PISA (2006)	All participating countries (many languages); Focus on UK (English) and Denmark (Danish)	Reading	Investigate methodological issues (model fit and DIF) in PISA	Andersen's (1973) Conditional Likelihood Ratio (CLR); Rasch model Item purification (Deletion of DIF items)	Data do not fit the Rasch model DIF is detected in a number of items Country ranking is not reliable
Huang et al. (2014)	PISA (2006)	USA (English); Canada (English); Hong Kong (Chinese); mainland China (Chinese)	Science	Investigate cross-language, cross-cultural validity across country pairs (3 variables: language, culture, curriculum coverage)	Multidimensional random coefficient multinomial logit model (MCRMLM) Rating of DIF items by a bilingual person and a panel of raters composed of teachers and students	Curriculum coverage is the most serious source of DIF Differential context familiarity is another significant source of DIF

Several of these studies indicated that test translation and adaptation did impact item difficulty (Ercikan & Koh, 2005; Grisay & Monsieur, 2007; Grisay et al, 2007; Oliveri & Ercikan, 2011; Kreiner & Christensen, 2014; Huang et al., 2014) and that translation and adaptation of constructs is not always successful (Ercikan & Koh, 2005; Hauger & Sireci, 2008). Oliveri and Ercikan (2011) found items with DIF when comparing English and French versions of the same PISA 2003 problem solving items in Canada. However at test level, no overall difference was conferred because of language differences. In this chapter, DIF analysis was carried out and potential advantages/disadvantages were evaluated at item level and at test level to explore whether DIF cancels out at test level.

Particularly relevant to this thesis, Grisay and Monseur (2007) found that the amount of DIF increased with distance between languages. For instance, there is less DIF between the French version used in Canada and the one used in France than there is DIF between the English and French versions used in Canada. Also, there is less DIF between ‘cousin languages’, for instance Spanish and French, than there is between distant languages, for instance Arabic and French.

Similar findings were reached in Huang et al.’s (2014) study of DIF between the English and Chinese versions of PISA science 2006 tests in four countries USA, Canada, Hong Kong and mainland China. Huang and her colleagues found that DIF was minimal between USA and Canada while it was much larger between USA and mainland China. It is however important to note that Huang et al. (2014) found large DIF between Hong Kong and mainland China although both countries have a common language. Huang et al. (2014) attributed the DIF to differential curricular coverage.

Moreover, Grisay and Monseur's (2007) study suggested that the amount of DIF increased as the gross domestic production (GDP) per country decreases. This finding is relevant to this research. A comparison between pairs of countries (UK-France; UK-Jordan; France-Jordan) can test this hypothesis. UK and France are both OECD countries, with 'cousin languages' and high GDP (respectively 2.435 trillion USD and 2.613 trillion USD in 2012). Whereas Jordan is not an OECD country, has a language that is distant to both English and French and has a much lower GDP (31.24 billion USD in 2012).

In this chapter, not only do I investigate whether a country is particularly dis/advantaged in the PISA Science 2006 tests and identify DIF items, I also investigate whether the distance between languages and the difference in GDP can be related to the direction of overall DIF. It is worth noting that, given that DIF is a relative measure, if most items favoured participants from high GDP countries, DIF would not be manifested. The next section describes the DIF approach adopted in this research.

7.2 DIF approach

In the current study, DIF analysis was carried out to investigate the impact of translation on item difficulty. The approach adopted was based on Hagquist & Andrich's (2004) method: a two-way ANOVA of response residuals (Hagquist & Andrich, 2004; Tennant & Pallant, 2007) of two factors ('country' and 'class interval') and their interaction. This approach has not been used previously in this type of analyses but it has the advantage of assessing general misfit, uniform DIF, and non-uniform DIF simultaneously. The following describes this approach.

7.2.1 DIF detection using ANOVA of response residuals

Response residuals were described in Chapter 6 (Section 6.5.2a, Formula 6.6). They (Z_{ni}) can be expressed as:

$$Z_{ni} = \frac{x_{ni} - E(x_{ni})}{\sqrt{V(x_{ni})}} \quad (7.2)$$

where x_{ni} is the observed response of person n to item i , $E(x_{ni})$ is the corresponding expected response with model variance $V(x_{ni})$.

Each person was assigned to a factor group, in this case a specific country (UK, France or Jordan), and a class interval based on his/her location on the latent trait. In this case, the latent trait was a given ability in science. Given that the focus was on the general trend of the functioning of an item across a country, the number of class intervals was taken as 4. This choice did not jeopardise having enough information per class interval, i.e. at least 30 persons per class interval (Hagquist & Andrich, 2004). For each item, response residuals were analysed with a two-way ANOVA and the interaction between person factor (i.e. country) and class factor (i.e. class interval) was investigated.

The presence of DIF was indicated by statistically significant inter-person-group variance. It was therefore possible to study the main effect of the country and the interaction between the class intervals and country. Swaminathan and Rogers (1990) described two types of DIF: uniform and non-uniform DIF. Uniform DIF occurs when an item behaves differently in the different countries in a consistent manner across class intervals. In this case, the item characteristic curves (ICCs) (See Chapter 6, Section 6.5.2c) of the three countries would be roughly parallel to each other. Non-uniform DIF was manifested when items behaved

differently in the different countries and inconsistently across class intervals. In the latter case, intersection of the ICCs of at least two countries is expected.

Uniform and non-uniform DIF were investigated using the same sample of 12,000 persons with equal proportions of students per country used for the Rasch analyses (Chapter 6, Section 6.2.3). This is consistent with the literature which recommends that the sample sizes of the groups considered should be as much as possible the same to avoid having DIF weighted by the estimates of the group with the larger sample (Andrich & Hagquist, 2012). To limit the probability of Type I error, the chi-square value was again adjusted to a sample of 1,635 and Bonferroni correction was applied with a base of 0.05 (i.e. $p < 0.000485$). The rationale behind adjusting the chi square value to a smaller sample and the Bonferroni correction was detailed in Chapter 6 (Section 6.3.5).

7.2.2 Detecting DIF: Artificial versus Real DIF

Procedures of investigating DIF induces an artefact manifested as additional DIF. Andrich and Hagquist (2012) named this type of DIF as *artificial DIF* to distinguish it from *real DIF* inherent in the items. The authors explained the way artificial DIF is induced in dichotomous items using the Mantel-Haenszel (1959) heuristic method as developed by Holland and Thayer (1988) as well as using the Rasch model (1961). Later, the authors generalise the concept of artificial and real DIF to polytomous items (Andrich & Hagquist, 2015).

In both cases, investigating DIF necessitates classifying people into class intervals. Given that the actual person locations on the continuum are not available, person estimates are used

instead. In the Rasch model, the person total score is a sufficient statistic (Andersen, 1977) and is used as a person estimate on which conditioning occurs to create class intervals.

For a test with three dichotomous items, the probability of a correct response on item i given the total score r can be expressed as $\pi_{ri} = \Pr\{X_{ni} = 1 \mid r\}$ (7.3)

With Item 2 correct for instance, the probability of getting a total score of r can be written as follows:

For $r = 1$, $\pi_{12} = \Pr\{(0,1,0) \mid r = 1\}$

For $r = 2$, $\pi_{22} = \Pr\{(1,1,0) \mid r = 2\} + \Pr\{(0,1,1) \mid r = 2\}$

For $r = 3$, $\pi_{32} = \Pr\{(1,1,1) \mid r = 3\}$

By definition, the sum of probabilities of correct responses on item i given the total score r is equal to the total score (Andrich & Hagquist, 2012) i.e. $\sum_{i=1}^I \pi_{ri} = r$. This constraint is central to artificial DIF. When an item shows real DIF its score for a class interval will be higher. Given that the total score across items is fixed, this constrains other items to have lower values and hence induces DIF to favour other groups in other items. This form of DIF is artificial because it results from the procedure of detecting real DIF and does not reflect real differences in the behaviour of an item across groups. Artificial DIF may obstruct the identification of items having real DIF. The direction of the bias, the location and the amount of artificial DIF is not predictable yet it is assumed to be uniformly spread over all the items (Andrich & Hagquist, 2012, 2015). The artificial DIF, induced by real DIF in an item and favouring the other groups, is spread across all other items and is the same in magnitude as real DIF.

Because of the presence of artificial DIF, and potential difficulty in distinguishing it from real DIF, Andrich and Hagquist (2012) have explained a sequential procedure to handle it with dichotomous items. The procedure consisted of treating one DIF item at a time starting from the one with the largest magnitude of DIF. This procedure is based on the logic of the inducement of artificial DIF and was recently extended to polytomous items (Andrich & Hagquist, 2015). The item with the largest DIF (i.e. highest mean square) is split into a number of items depending on the number of groups.

In this research, the item with the largest DIF was split into three items based on the three groups; i.e. countries (UK, France and Jordan). The original item is eliminated in the response matrix and replaced by the three split items which have either people from UK only or from France only or from Jordan only responding to them. The new items which are country-specific cannot be responsible for inducing any artificial DIF in other items. If the split item is the only item with real DIF and was forcing other items to show artificial DIF then when split, artificial DIF would disappear. However, if more items showed DIF after splitting the one with the highest amount of DIF, then the same procedure could be repeated until no items showed DIF.

The analysis of DIF is not thoroughly described in the technical reports for PISA 2006 items and the way DIF is handled is unclear. Potential item-by-country interactions were analysed using parameter estimation of IRT models for each country separately then results (namely the discrimination parameter estimates) were compared with those of the international sample (OECD, 2009b). A full set of item-by-country interaction plots was constructed for cognitive items and interest (embedded) items. Results suggest that ‘the embedded item parameter estimates are more stable across countries than the parameter estimates for the cognitive

items' (OECD, 2009b, p.356). Kreiner and Christensen (2014) used a similar procedure to the one described by Andrich and Hagquist (2012) to eliminate artificial DIF when analysing measurement bias in PISA reading 2006 data. In this research, artificial DIF is eliminated using Andrich and Hagquist's (2012) methods and real DIF was identified. The next section describes the results of DIF analyses after eliminating artificial DIF.

7.3 Results

The following section presents the results of the DIF analyses at item level and at test level. The PCA carried out in Chapter 6 suggests that at test level, DIF (at item level) balances, and does not seem to confer any relative advantages or disadvantages to any particular country. However, at item level, DIF exhibited by items shows lack of invariance and this can provide important insight and contribute to a better understanding of the variable examined.

7.3.1 DIF at item level

After splitting items using Andrich and Hagquist's (2012, 2015) method (thus eliminating artificial DIF), and with a Bonferroni correction base of 0.05 (i.e. at the probability/total number of items = $0.05/103 = 0.000485$ level) and a chi square value adjusted to a sample of 1,635, the DIF analysis identified 23 out of 103 items (that is, 22%) with real DIF.

Results of the two-way ANOVA of residuals from four class intervals of these 23 items are presented in Table 7.2 below. Column 1 of Table 7.2 summarises the mean square (MS), the F statistic and probability values for the country main effect and the class interval by country effect of the 23 out of 103 items that had a main effect (significant probabilities >0.000485).

Note that RUMM2030 (Andrich et al., 2009) provides the MS, F statistics and probability values for the class interval main effect. However, the results are similar to the test of fit in Table 6.2 (Chapter 6). Since this is not the focus of this analysis, these results were not presented in Table 7.2.

Results in Table 7.2 suggest that the 23 items behaved consistently differently in at least two countries across all class intervals; this is evidence of uniform DIF. No items showed any significant interaction between country and class interval meaning that non-uniform DIF was not detected.

Table 7.2: Two-way ANOVA of residuals from four class intervals

Item	Country				Country X Class Interval			
	MS	F	df	P	MS	F	df	P
1	87.63200	18.283	2	0.000000	-7.6213	-1.59009	6	0.99999
9	88.64477	15.20257	2	0.000005	15.01624	2.57528	6	0.01824
15	85.22733	13.04631	2	0.000013	-0.64196	-0.09827	6	0.99999
18	283.2216	36.00194	2	0.000003	-18.0498	-2.29355	6	0.99999
23	253.5416	37.26409	2	0.000008	9.55278	1.40401	6	0.21115
30	88.16107	11.17145	2	0.000016	-3.1102	-0.39411	6	0.99999
41	97.34167	10.84036	2	0.000020	2.66617	0.29691	6	0.93833
43	59.56032	8.38595	2	0.000255	7.72309	1.08739	6	0.36882
50	85.51193	11.68901	2	0.000008	2.31994	0.31712	6	0.92806
51	61.20754	8.38158	2	0.000262	5.81376	0.79612	6	0.57327
54	86.40760	13.25297	2	0.000009	2.87138	0.4404	6	0.85177
58	209.1526	29.95152	2	0.000006	11.0838	1.58725	6	0.14882
59	83.03282	14.40456	2	0.000001	-1.46531	-0.2542	6	0.99999
62	54.52460	8.63718	2	0.000217	6.44401	1.02079	6	0.41083
63	78.11069	10.68339	2	0.000018	0.24319	0.03326	6	0.99985
67	143.5584	18.51765	2	0.000001	3.61631	0.46647	6	0.83322
76	120.3432	17.22393	2	0.000000	11.68403	1.67226	6	0.12587
86	72.20386	9.81187	2	0.000054	4.26000	0.5789	6	0.74726
90	69.52485	9.85556	2	0.000066	7.82726	1.10956	6	0.35552
94	75.94308	12.78919	2	0.000005	4.58175	0.77159	6	0.59252
95	96.57841	11.75653	2	0.000008	-2.38886	-0.2908	6	0.999999
96	88.08359	9.85296	2	0.000064	4.86414	0.5441	6	0.774654
103	53.79768	8.18277	2	0.000309	4.98653	0.75847	6	0.602918

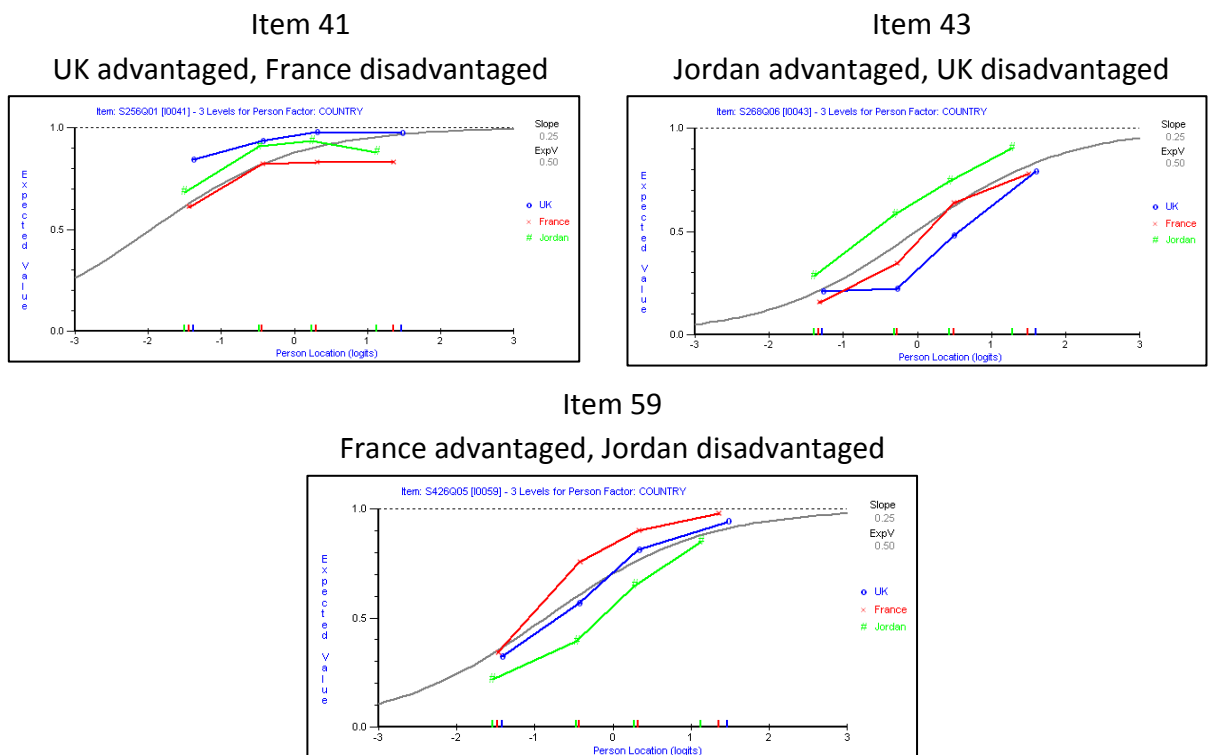
MS: mean square, F: F-statistic; df: degrees of freedom; P: Probability

Uniform DIF was investigated graphically using ICCs. Figure 7.1 below shows three examples of DIF items, items 41, 43 and, 59 with different countries being advantaged or disadvantaged. The ICCs of the remaining items showing DIF can be found in Appendix 7.1.

As shown in Figure 7.1 below, all DIF items show uniform DIF: the observed proportions of correct responses around the ICCs of the countries are ‘parallel’ and do not intersect. Instances of intersection such as the intersection between the lines connecting the observed proportions of France and UK in item 43 were not flagged as significant in Table 7.2. ‘Parallel’ lines

connecting the observed proportions indicate that the behaviour of an item is consistent across class intervals within a given country and that there was no interaction between country and class interval. For instance, item 41 advantages UK and disadvantages France with Jordan being in between. This observation is true across all class intervals. Similar observations could be made for items 43 and 59. Item 43 advantages Jordan and disadvantages UK and item 59 advantages France and disadvantages Jordan.

Figure 7.1: Examples of DIF items (UK: blue, France: red, Jordan: green)

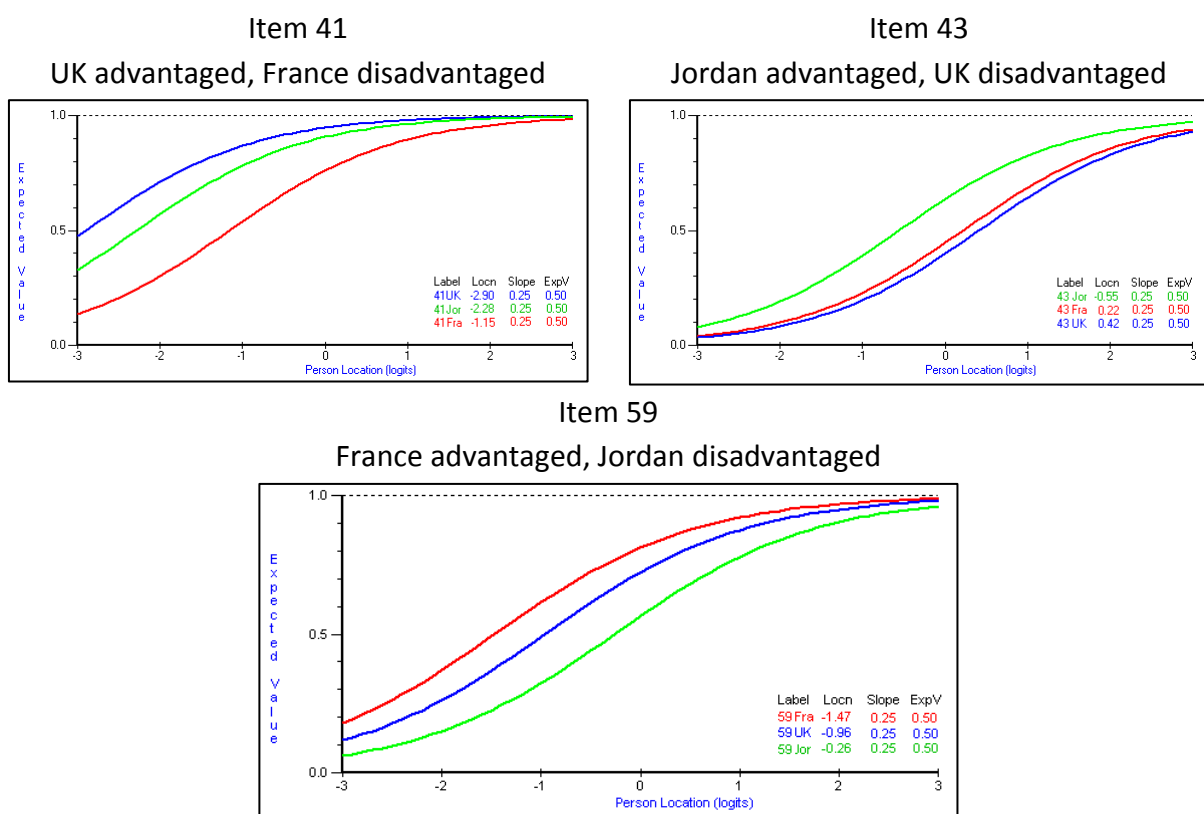


7.3.2 Quantifying DIF at item level

The relative item location for each country, obtained by splitting the responses of the original item by country, can be visualised by observing the country ICCs for every item. Figure 7.2 represents the country ICCs for the same items illustrated in Figure 7.1.

The country ICCs are consistent with findings reported earlier on these items (Figure 7.1). Item 41 is easy for all countries but seems to be easier for UK (item location -2.90) than it is for France (item location -2.28) or Jordan (item location -1.15). Item 43 appears to be harder for UK (item location 0.42) and France (item location 0.22) and easier for Jordan (item location -0.55). Item 59 seems to be harder for Jordan (item location -0.26) and easier for UK (item location -1.47).

Figure 7.2: Country ICC for items 41, 43 and 59



The presence of DIF in 22% of the items raises concerns about bias at test level. Is any of the countries particularly advantaged or disadvantaged overall by the test? This question is investigated in the next section.

7.3.3 Was a particular country dis/advantaged overall?

This section examines whether DIF items consistently favoured or unfairly challenged a particular country in PISA 2006 science tests. Grisay and Monseur's (2007) study indicated that countries with 'cousin languages' and with similar GDP indices have fewer items showing DIF when compared together than countries with disparate languages and different GDPs. Besides, Grisay et al. (2007) highlighted the low level of equivalence of Arabic versions with the PISA source versions. Based on this, one can hypothesise that since Jordan has a very different language and has a significantly lower GDP than the other two countries (see section 5.7), it would be disadvantaged while UK and France, having similar languages and GDP, may have insignificant differences when compared to each other.

Advantaged and disadvantaged countries were determined visually for every DIF item by examining the country ICCs. Results are summarised in Table 7.3. Based on Table 7.3, advantages and disadvantages are not consistently conferred onto a specific country, and they are not evenly distributed either. Unlike the predictions I mentioned earlier, France seems to be the most disadvantaged with the highest number of items showing DIF (10 as opposed to 6 in the UK and 7 in Jordan) and the lowest number of advantaging items (5 as opposed to 7 in the UK and 11 in Jordan). Jordan seems to be the most advantaged with the highest number of advantaging items and a similar number of disadvantaging items as the UK

Table 7.3: Number of relatively dis/advantaging DIF items per country

	UK	FRA	JOR	Total
ADV	7	5	11	23
DISADV	6	10	7	23

It is important to stress that not all dis/advantages presented in Table 7.3 are statistically significant. It is very possible that the relative advantages in one country, even if conferred by a lower number of items, outweigh the relative disadvantages resulting from a larger number of items. However, another possible hypothesis based on Table 7.3 is that France is disadvantaged compared to the other two countries. Next, I investigate these hypotheses by first comparing the means of DIF item locations using a repeated-measures ANOVA; then comparing the person means before and after splitting.

7.3.3.a Comparing means of DIF item location

The average difficulty of the DIF items per country was computed by summing up all DIF item locations for every country and dividing by the total number of DIF items; i.e. dividing by 23. Table 7.4 presents DIF item locations and the average item difficulty per country. Table 7.4 suggests that, overall, DIF items are easiest for Jordanian examinees (average difficulty of DIF items = -0.22) and most difficult for French participants (average difficulty of DIF items = 0.19) with the UK being in between with an average difficulty of DIF items being - 0.15. Hence, it looks like the second hypothesis, where France is advantaged, as a more plausible one.

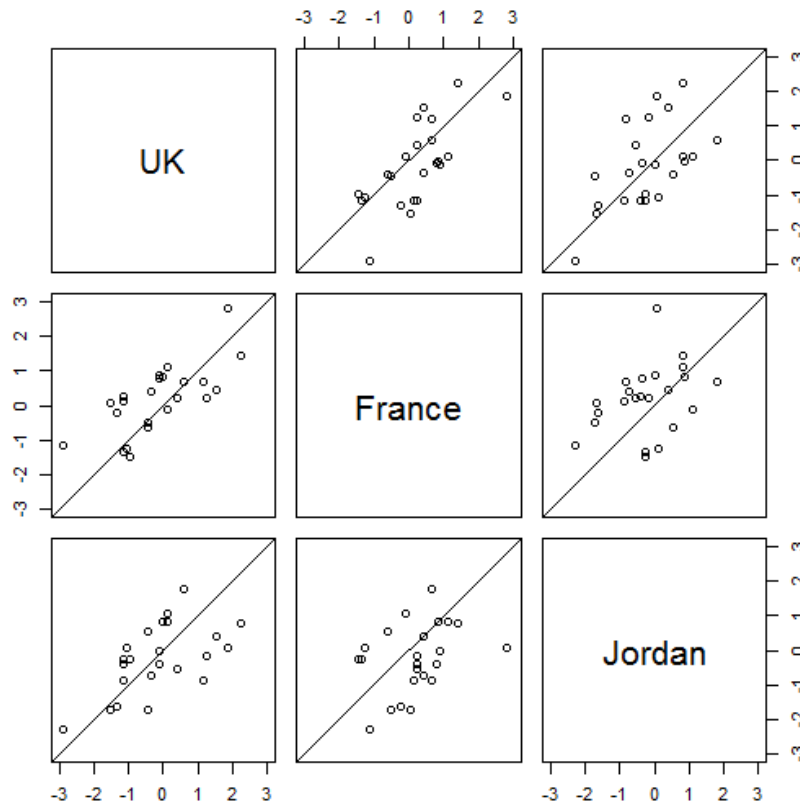
Table 7.4: DIF item locations and average difficulty per country

Item	Item Location		
	UK	France	Jordan
1	-0.43	-0.62	0.53
9	1.26	0.23	-0.15
15	0.59	0.67	1.80
18	1.17	0.67	-0.84
23	1.84	2.81	0.07
30	2.24	1.43	0.80
41	-2.90	-1.15	-2.28
43	0.42	0.22	-0.55
50	1.53	0.43	0.40
51	0.12	1.13	0.84
54	-1.15	-1.34	-0.26
58	-1.53	0.07	-1.69
59	-0.96	-1.47	-0.26
62	-1.17	0.25	-0.39
63	-0.46	-0.50	-1.70
67	-1.17	0.13	-0.85
76	-1.06	-1.25	0.10
86	-0.04	0.84	0.85
90	0.10	-0.11	1.09
94	-1.32	-0.22	-1.60
95	-0.10	0.78	-0.37
96	-0.13	0.87	0.00
103	-0.34	0.41	-0.71
Sum of difficulties	-3.49	4.28	-5.17
Average difficulty	-0.15	0.19	-0.22

A scatter plot of pairs of countries was constructed based on locations of DIF items (see Figure 7.3 below). The scatter plot paints a different picture. UK and France were more similar than did UK and Jordan or France and Jordan. This is not consistent with previous results of Table 7.4 which suggests that the distance between mean locations of DIF items of UK and Jordan is the least ($-0.15 - (-0.22) = 0.07$ logits) followed by the distance between

mean locations of DIF items of UK and France ($-0.15 - 0.19 = -0.34$ i.e. 0.34 in absolute value) and then France and Jordan ($0.19 - (-0.22) = 0.41$).

Figure 7.3: Scatter plot matrix of pairs of countries based on locations of DIF items.



These results convey Grisay and Monseur's (2007) observations about proximity of languages and GDP. UK and France are more similar to each other than they are to Jordan. These conclusions are not based on significance tests. To test whether the means of the locations of split items were significantly different at 5% significance level, a repeated-measures ANOVA was carried out. A repeated-measures ANOVA was considered rather than a one-way ANOVA because there is dependence amongst the country-specific item locations in every row: they all relate to the same item and are all constrained by the same overall item location. Within-subjects effects are displayed in Table 7.5 below.

The condition of sphericity of variances was met (Mauchly's test statistic = 0.92; probability = 0.39 > $p = 0.05$). Results of Table 7.5 indicated that the within-subject effects were not significant ($F = 2.137$; $p = 0.130 > 0.05$). This suggested that there was no significant net advantage or disadvantage given to any country. The advantages and disadvantages conferred by specific items were balanced out throughout the exam. These results support the robustness of the comparisons across countries from a technical point of view.

Table 7.5: Within-subjects effects (repeated-measures ANOVA) at 5% significance level

Source	Sum of Squares	df	Mean Square	F	Significance
Item location	2.210	2	1.105	2.137	0.130
Error (Item location)	22.753	44	0.517		

Despite having an insignificant repeated-measures ANOVA, pairwise comparisons were explored to investigate how countries compared with each other. Results are displayed in Table 7.6 below.

Table 7.6: Pairwise comparisons

Contrast	Value of Contrast	Standard Error	Significance
UK-FRA	0.338	0.181	0.227
UK-JOR	-0.073	0.216	1.000
FRA-JOR	-0.411	0.235	0.285

^a Bonferroni adjustment for multiple comparisons

The results in Table 7.6 suggest that no contrast was significant (all probabilities > 0.05) confirming that means were not significantly different. At test level, no country was advantaged or disadvantaged significantly.

7.3.3.b Variation between person means before and after splitting

Another way of examining whether a particular country is advantaged or disadvantaged by DIF items is by comparing the person means before and after splitting. If for any particular

country, the splitting of items resulted in an increase of the average person mean of a particular country then this suggests that the DIF items were on average more difficult for that country. However, if splitting led to the decrease of the person mean of a given country, then this indicates that on average, DIF items were easier for this country. This approach has not been carried out in any of the studies reviewed investigating DIF in international assessments.

The estimates of person means before and after splitting cannot be directly compared because each analysis has its own origin – the item estimates in both analyses sum to zero, even though there is effectively a different number of items in the two analyses. To establish a comparison using a common origin, an *average item anchoring* analysis was carried using the ‘Average Item Anchoring’ function in RUMM2030 (Andrich et al., 2009). In the analysis in which DIF items are split, the mean of non-split items was fixed to the value of their estimates in the analysis where no items were split. The parameters of all items were reestimated. The two analyses effectively had the same origin, and allowed direct comparison of person means generated from the two analyses. Results of the person means before and after splitting were displayed in Table 7.7 below.

Table 7.7: Person means before and after splitting with fixed mean

CNT	Sample	Before splitting ¹		After splitting		D Av	SD(D Av)	SD(D)	t	ES
		M	SD	M	SD					
UK	4000	0.38	1.16	0.36	1.2	-0.01696	0.00078	0.04939	-439.6	-0.3433
FRA	4000	0.21	1.15	0.26	1.16	0.049794	0.00066	0.04186	1797.6	1.1897
JOR	4000	-0.49	1.00	-0.53	1.02	-0.04035	0.00076	0.04782	-1116.2	-0.8439

CNT: country; M: mean; SD: standard deviation; D: deviation; D Av: average deviation; ES: effect size

¹ origin fixed;

Based on Table 7.7, the deviations of person means before and after splitting are negative in UK and Jordan and positive in France. This suggests that the DIF items were on average

easier for participants in UK and Jordan while they were harder for students in France. Hence, participants in UK and Jordan were relatively advantaged by the DIF items while their counterparts in France were relatively disadvantaged. The decrease in person means of participants from UK and Jordan indicates that after splitting, participants from UK and Jordan received less credit for getting DIF items correct because they found them relatively easy. The opposite is true for their counterparts in France. The increase in person mean suggests that they have received more credit for getting DIF items correct because they considered them harder.

Statistical significance of the deviations was carried out using two-tailed t tests for dependent samples. All three t values are highly significant at 0.05. Hence at 5% confidence level two-tailed, the differences between person means before and after splitting items are significantly different for all three countries. This is not surprising given the large sample size and the very small standard deviation of the differences.

Although statistically significant, the actual magnitude varies across the countries based on Cohen's (1962) classification of effect sizes with $ES \sim 0.20$ indicating a small effect, $ES \sim 0.50$ indicating a medium effect and $ES \sim 0.80$ indicating a large effect. The difference between the means before and after splitting items in the UK seems to have a relatively low effect size (0.3 in absolute value) while it is of a greater effect size for France (1.2) and Jordan (0.84 in absolute value). It is worth noting however, that the values of ES are inflated because of the very small values of the standard deviations of the differences. Hence, DIF does significantly impact person estimates on average in France and Jordan in opposite directions while it mildly does in UK. Nevertheless, the magnitude of the average deviation in person

estimates in relatively small (less than 0.1 logits) for all of the items. This negligible effect is consistent with the repeated-measures ANOVA study discussed in Section 7.3.3.b.

7.3.4 Conclusion of the DIF analyses

In this section, results of a DIF analysis conducted as a two-way ANOVA of response residuals were presented. After eliminating artificial DIF induced by procedures of detecting real DIF using Andrich and Hagquist's (2012; 2015) method, 23 items showed significant DIF. A repeated-measures ANOVA of locations of DIF items suggested that, despite nearly a quarter of items being flagged for DIF, means of DIF items were not significantly different. This result was corroborated by the significant yet negligible deviations of person means after splitting items from their initial values before splitting for each country. DIF seems to balance out making tests invariant at test level. These findings were consistent with the results of the PCA carried out in chapter 6 which suggested that the tests were unidimensional at test level.

The focus of this chapter was to investigate whether invariance was held at item level. A complementary approach to DIF analyses is a distractor analysis of multiple choice items. Proportions of students selecting the key (i.e. the correct answer) or any of the distractors (i.e. the incorrect alternatives) in a multiple choice question may vary from one country to another. The key or the distractors could be more or less attractive for students to select in a particular country because of the way the key or the distractor is formulated. The key may include a subtle unintentional hint to the correct response that may increase its probability of being selected. Alternatively, the key could be formulated in a confusing manner decreasing its probability of being selected. A similar yet reverse logic is true for distractor selection. In the

PISA science 2006 test, 64 items were multiple choice items (items 39 to 102). The following section describes the distractor analysis carried out and summarises its results.

7.4 Distractor Analysis

Distractors are the wrong alternatives provided to students along with the key (i.e. the right choice) in multiple choice items. Distractors often reflect frequent mistakes and common misconceptions examinees have. For this reason, they could be informative. Herrman-Abell and DeBoer (2011) used distractor analyses along with the Rasch model on data from a test administered to a large sample of students in the USA to investigate misconceptions in chemistry and detect structural problems with items. Distractor analyses often complement initial investigations and allow the researcher to have access to more information about the way a particular item behaves in different groups. Sadler (1998) combined distractor analyses with qualitative item analyses to construct models of students' science conceptions. Hauger and Sireci (2008) used it to complement their DIF analyses of TIMSS items between native and non-native speakers in Singapore, USA and Iran.

In this thesis, a distractor analysis may provide evidence to several empirical questions. To what extent are distractors faithfully translated and adapted across languages? Are they made more misleading and confusing in different languages? Or alternatively, do they provide more clues in one language than they do in another? Distractor analyses were carried out on a subset of the multiple choice items (i.e. items 39 to 102) irrespective of whether items were flagged for DIF or not. The rationale behind casting a wider net is to tease out more information about the way the key and distractors behaved in the three countries of interest regardless of whether

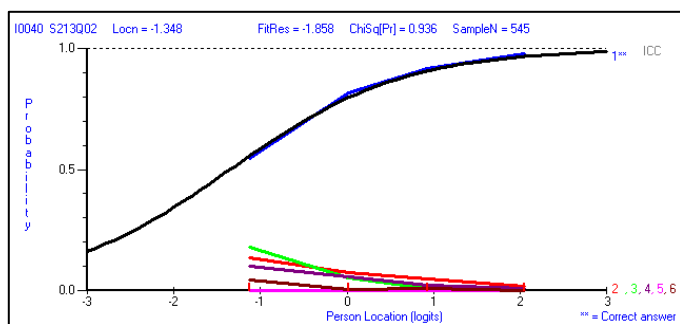
on the whole the item showed statistically significant DIF. This section describes the approach and provides the results of the analysis.

7.4.1 Distractor proportion curves (DPCs)

RUMM2030's (Andrich et al., 2009) 'MC Distractor Curve' function was used to obtain distractor proportion curves (DPC) for each item. DPCs are graphical representations of the empirical proportions of each alternative relative to the ICC across class intervals. DPCs help investigate whether the key fit the model well and whether any of the incorrect alternatives had a higher selection rate than expected.

Figure 7.4 below provides an example of the output: the DPCs of item 40 in the UK. The key in the item in Figure 7.4 is option 1 (blue). As shown in the figure, the observed data of that option match the ICC of the item. The proportion of selecting this option increases with person ability as expected by the model. Meanwhile, the proportions of the other options, distractors 2, 3, 4 are low and decrease with person ability. It is important to note that in this dataset the curves labelled with the highest digit do not represent distractors. They are a product of the re-coding of the data (see Appendix to 6.1) and represent the proportions of 'multiple responses' on a given item and the 'embedded missing responses' in the item (See chapter 6, Section 6.2.1). For instance, in Figure 7.4 below, curve 5 represents the proportion of multiple responses to item 40 and curve 6 represents the proportion of embedded missing responses. These curves are not discussed further as they do not provide information about the behaviour of distractors and they are close to zero in all items for all three countries.

Figure 7.4: DPCs of Item 40 from the UK sample



7.4.2 Data

The PISA science 2006 test consists of 103 cognitive items out of which 64 items are multiple choice questions (MCQs); these being either simple or complex. Simple MCQs consist of a statement and 3 or 4 alternatives, one of which is the key (that is, the correct response) and the rest being distractors (that is, incorrect options). Figure 7.5a is an example of such an item²⁰.

Figure 7.5a: Simple MCQ (PISA Science 2006)

Question 2: CLOTHES

S213Q02

Which piece of laboratory equipment would be among the equipment you would need to check that the fabric is conducting electricity?

- A Voltmeter
- B Light box
- C Micrometer
- D Sound meter

CLOTHES SCORING 2

Full Credit

Code 1: A. Voltmeter.

No Credit

Code 0: Other responses.

Code 9: Missing.

²⁰ See the following website for released PISA items and scoring: www.oecd.org/pisa/38709385.pdf.

Complex MCQs consist of a statement with 3 or 4 statements to which the examinee should respond with YES or NO. The examinee is only granted a full credit when s/he answers all statements correctly. Figure 7.5b an example of such an item²¹.

Figure 7.5b: Complex MCQ (PISA Science 2006)

Question 1: CLOTHES

S213Q01

Can these claims made in the article be tested through scientific investigation in the laboratory?

Circle either "Yes" or "No" for each.

The material can be	Can the claim be tested through scientific investigation in the laboratory?
washed without being damaged.	Yes / No
wrapped around objects without being damaged.	Yes / No
scrunched up without being damaged.	Yes / No
mass-produced cheaply.	Yes / No

CLOTHES SCORING 1

Full Credit

Code 1: Yes, Yes, Yes, No, in that order.

No Credit

Code 0: Other responses.

Code 9: Missing.

The distinction between simple and complex MCQs was only possible for released items because their detailed scoring was released too. Distractor analyses are only substantively meaningful when carried out on released simple MCQs. The DPCs yielded for complex MCQs do not represent distractors; they represent the number of correct responses. Amongst the released items, fifteen are MCQs out of which ten are simple MCQs. Three out of the ten released simple MCQs were flagged for DIF while seven were not. As mentioned earlier, the distractor analysis is carried out on both DIF and non-DIF simple MCQ items. Table 7.8 below lists the items of interest and divides them into DIF and non-DIF items.

²¹ See the following website for released PISA items and scoring: www.oecd.org/pisa/38709385.pdf.

Table 7.8: Items eligible for distractor analysis

Item number	PISA code (Unit)	DIF	non-DIF
40	S213Q02 (<i>Clothes</i>)		X
58	S426Q03 (<i>The Grand Canyon</i>)	X	
59	S426Q05 (<i>The Grand Canyon</i>)	X	
68	S447Q02 (<i>Sunscreens</i>)		X
69	S447Q03 (<i>Sunscreens</i>)		X
70	S447Q04 (<i>Sunscreens</i>)		X
80	S477Q02 (<i>Mary Montagu</i>)		X
81	S477Q03 (<i>Mary Montagu</i>)		X
85	S485Q03 (<i>Acid Rain</i>)		X
94	S508Q03 (<i>Genetically modified crops</i>)	X	

In order to compare the DPCs across countries, the initial sample ($n = 12,000$) was divided into the country samples: UK ($n_{UK} = 4,000$); France ($n_{FRA} = 4,000$) and Jordan ($n_{JOR} = 4,000$) using the function ‘Delete Sample by Person Factor’ available in RUMM2030. For instance, the UK sample was formed by deleting the French and Jordanian samples and so on. DPCs of all multiple choice items were generated for each country. The next section presents some of the results.

7.4.3 Results

A side by side comparison of DPCs of each of the ten released simple multiple choice items was carried out. Results are presented below. For the sake of a more organised presentation of results, the DPCs of the three items flagged for DIF are discussed first.

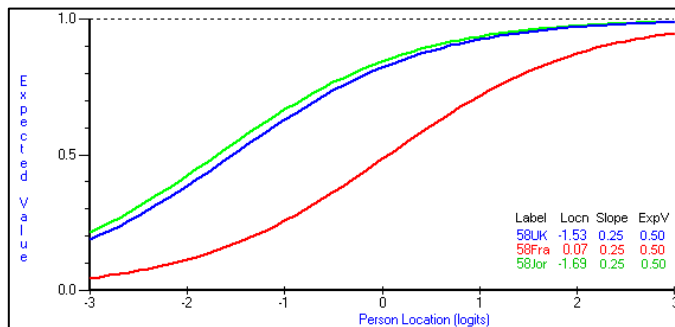
7.4.3.a Released simple MCQs flagged for DIF

Items 58, 59 and 94 are simple multiple choice items that were flagged for DIF. Each item is discussed at a time with a reference to the results of the DIF analysis.

Item 58 – S426Q03 (Grand Canyon)

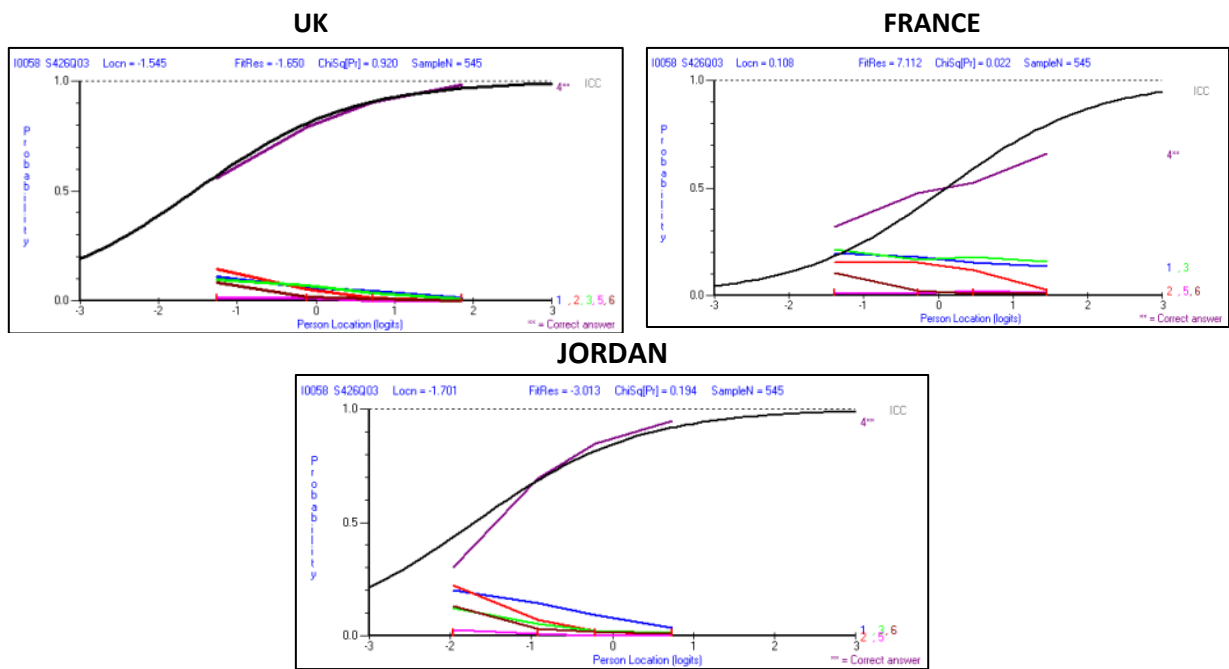
Item 58 was more difficult for the French (item location = 0.07) than it was for the UK (item location = -1.53) or Jordan (item location = -1.69) as shown in Figure 7.6.

Figure 7.6: Item locations per country for item 58



The results of the distractor analysis are displayed below in Figure 7.7. For all three countries the proportion of students choosing option 4, the key, is higher and increases with person ability while the proportions of participants selecting the other options, the distractors, are lower and decrease with person ability. For the UK, the key perfectly fits the ICC curve and the proportions of the distractors start from a low value and rapidly decrease with person ability. While the big picture is similar in Jordan, we notice that the proportion of students choosing the key increases more rapidly and exceeds the model expectations. The DPCs of France show a pattern of indecision amongst the alternatives. The DPC of the key does not fit the model; the curve is flatter than expected indicating relatively low discrimination and potential guessing. Distractors 1, 2 and 3 are almost equally attractive for low and medium ability students and proportions of distractors 1 and 3 never approach 0 even for very high ability students. This result is worth exploring further qualitatively – what is in distractors 1 and 3 attracting examinees in France?

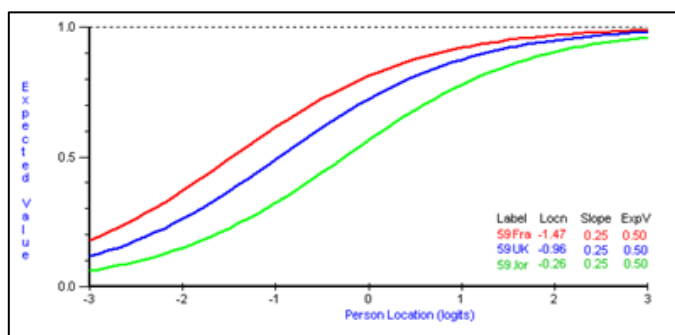
Figure 7.7: Distractor analysis of item 58



Item 59 – S426Q05 (Grand Canyon)

Item 59 was somehow more difficult for the Jordanian sample (item location = -0.26) than it was for the UK sample (item location = -0.96) or the French (item location = -1.47) as shown in Figure 7.8.

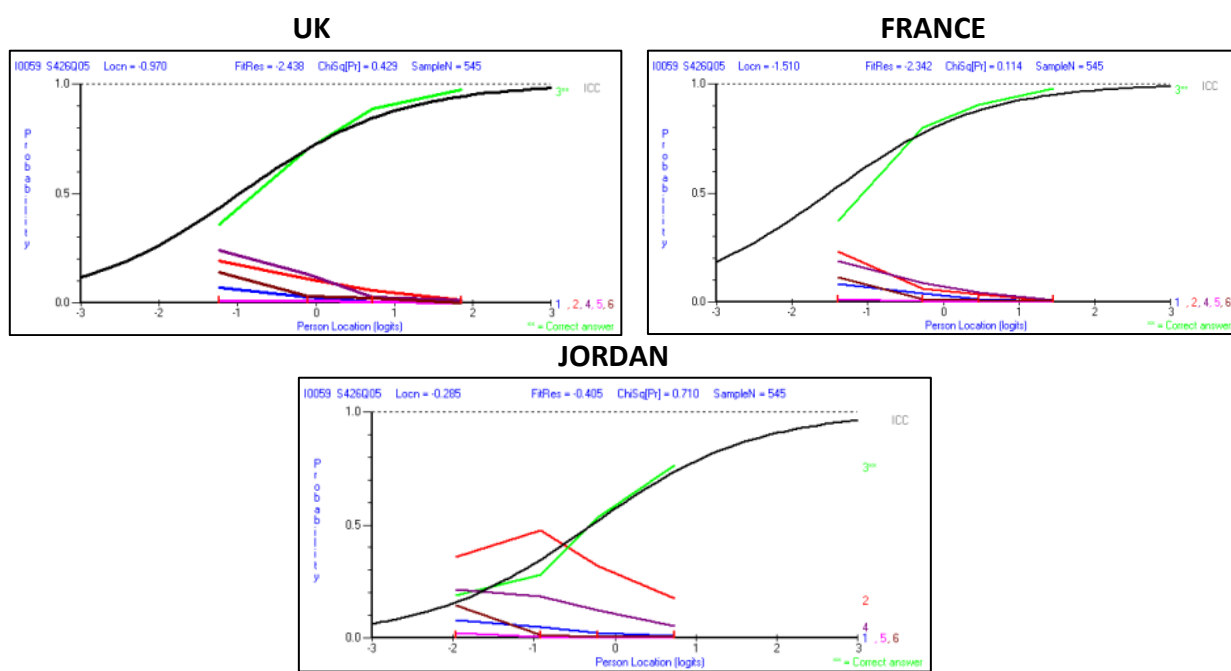
Figure 7.8: Item locations per country for item 59



The results of the distractor analysis are displayed below in Figure 7.9. For all three countries, the proportion of students selecting option 3, the key, is higher and increases with person

ability while the proportions of individuals choosing any of the other options, the distractors, decrease with person ability. For the UK and France, the key fits the ICC curve and the curves of the distractors start from a low value and rapidly decrease with person ability to become nil. Whereas in Jordan, we notice that the proportion of students choosing distractor 2 is higher than the key for low ability students and it remains relatively high (proportion = 20%) even for high ability students. Distractor 2 may be containing a misleading feature attracting Jordanian students, even those who are of high ability in science. This result is worth following up qualitatively – why does distractor 2 attract examinees in Jordan?

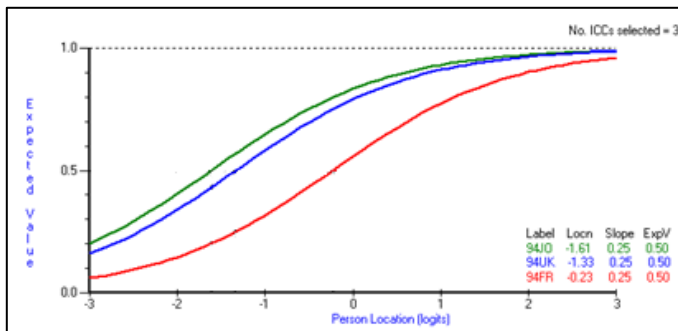
Figure 7.9: Distractor analysis of item 59



Item 94 – S508Q03 (Genetically Modified Crops)

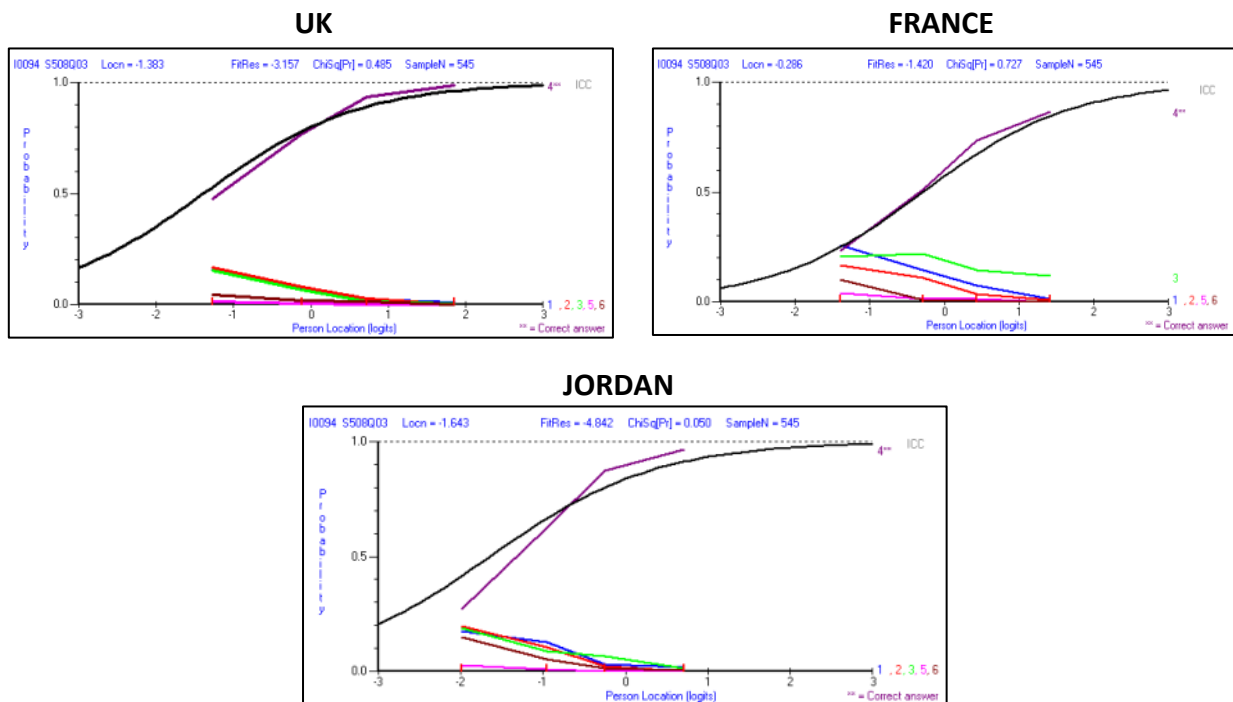
Item 94 was relatively more difficult for the French sample (item location = -0.23) than it was for the UK sample (item location = -1.33) or the Jordanian (item location = -1.61) as shown in Figure 7.10.

Figure 7.10: Item locations per country for item 94



The results of the distractor analysis are displayed below in Figure 7.11. For all three countries the proportion of the key, option 4, is higher and increases with person ability while the proportions of all the other options, the distractors, decrease with person ability. In the UK DPC, there is a very good fit between the key and the ICC curve and the proportions of students selecting the distractors start from a low value and rapidly decrease with person ability to become nil. A similar description could apply to the Jordanian DPC except that the proportion of individuals choosing the key increases more rapidly with person ability. The French DPC, points to a relatively high proportion of choosing option 3; this proportion remains above 10% for high ability students. It is worth investigating whether there is a qualitative explanation making the key easier to identify in Jordan and distractor 3 more attractive to French students.

Figure 7.11: Distractor analysis of item 94



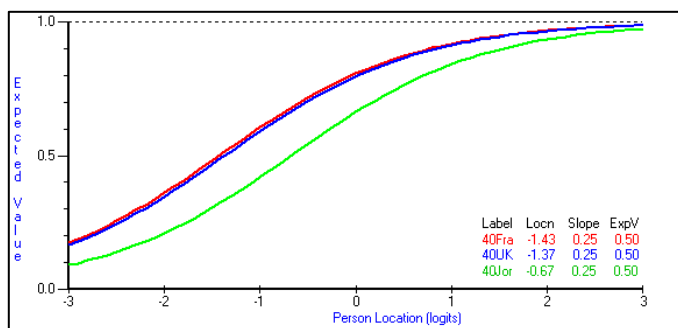
7.4.3.b Released simple MCQs not flagged for DIF

Items 40, 68, 69, 70, 80, 81 and 85 are simple multiple choice items that were not flagged for DIF. A distractor analysis may reveal information about differences in country behaviour including for items not detected for DIF. In the following, the distractor analysis of each item is discussed.

Item 40 – S213Q02 (Clothes)

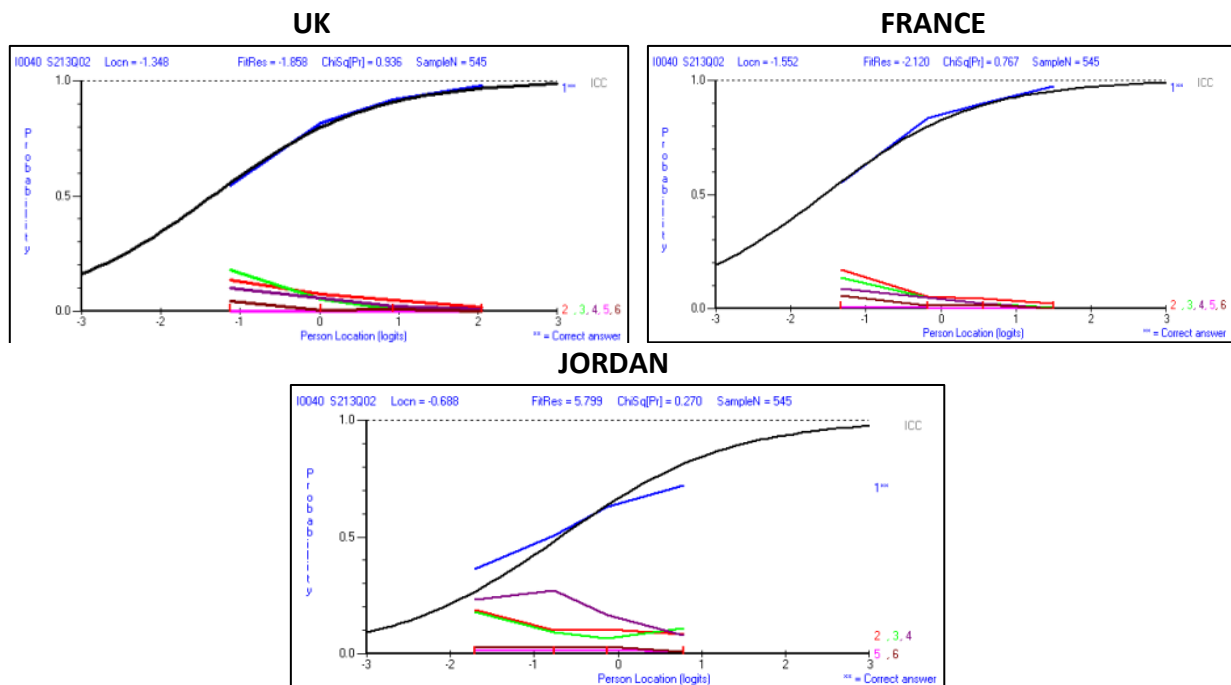
Item 40 was relatively more difficult for the Jordanians (item location = -0.67) than it was for the UK (item location = -1.37) or France (item location = -1.43) as shown in Figure 7.12 although overall DIF was not significant.

Figure 7.12: Item locations per country for item 40



The results of the distractor analysis are displayed below in Figure 7.13. For all three countries, the proportion of option 4, the key, is higher and increases with person ability while the proportions of all the other options, the distractors, are lower and decrease with person ability. For the UK and France the key perfectly fits the ICC curve and the proportions of the distractors start from a low value and rapidly decrease with person ability. However, in the Jordan DPC we notice that, although the proportion of individuals choosing the key increases with person ability, it remains lower than expected by the model for high ability students. Besides, the proportion of choosing the other distractors does not approach zero and distractor 4 is popular amongst low ability students. This result is worth exploring further qualitatively – what feature(s) in distractor 4 attracted examinees in Jordan?

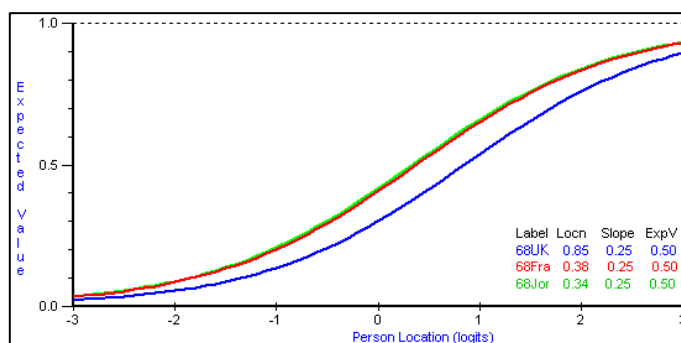
Figure 7.13: Distractor analysis of item 40



Item 68 – S447Q02 (Sunscreens)

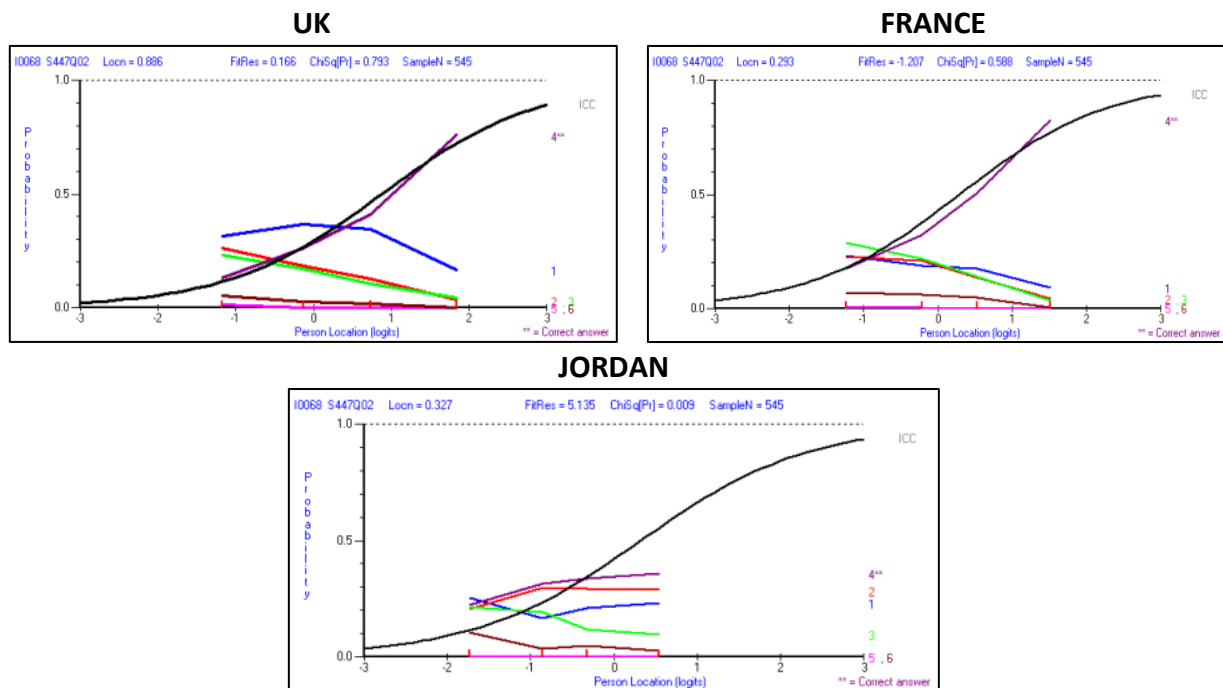
Although overall DIF is not significant, item 68 seemed more difficult for the UK sample (item location: 0.85) than it was for the Jordanians (item location: 0.34) or French (item location: 0.38) as shown in Figure 7.14.

Figure 7.14: Item locations per country for item 68



The results of the distractor analysis are displayed below in Figure 7.15.

Figure 7.15: Distractor analysis of item 68

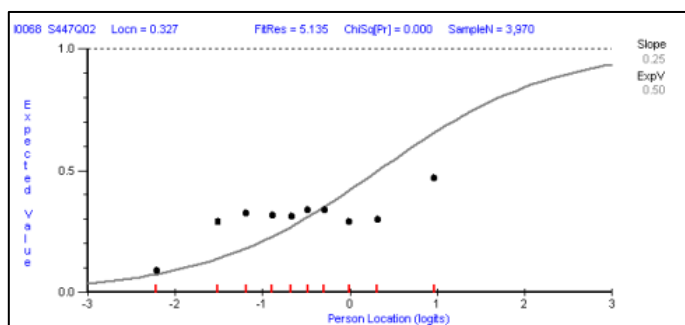


There is a clear difference in the DPCs across countries for this item. For France, the proportion of students selecting the key; i.e. option 4, is higher and increases with person ability while the proportions of all the other options, the distractors, are lower and decrease with person ability. This stands for the UK as well; however, the proportion of individuals choosing option 1 remains higher than expected by the model even for high ability students (proportion > 0.15). The Jordan DPC is a clear case of very low discrimination and random guessing. The key curve is flatter than the model despite it being the alternative with relatively the highest proportion. All distractors have a proportion of selection of 0.10 and above even for high ability students. Indeed, distractors 1 and 2 have a proportion of selection greater than 0.20.

Figure 7.16 below shows the ICC curve of item 68 for Jordan only. Evidence of random guessing can be detected with low ability students performing better than expected by the

model and high ability students performing worse than expected by the model. For a detailed discussion of random guessing refer to Chapter 6 (Section 6.6).

Figure 7.16: ICC curve of item 68 in Jordan

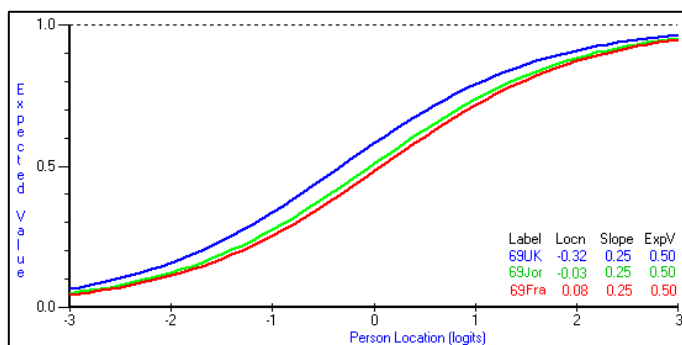


A more in-depth substantive analysis may reveal the potential reasons causing distractor 1 to be more attractive in the UK and explain why this item triggers random guessing in Jordan. One explanation of random guessing could be that it is because the item is amongst the 30 most difficult items. This remains a speculation and more qualitative evidence is required to test this hypothesis.

Item 69 – S447Q03 (Sunscreens)

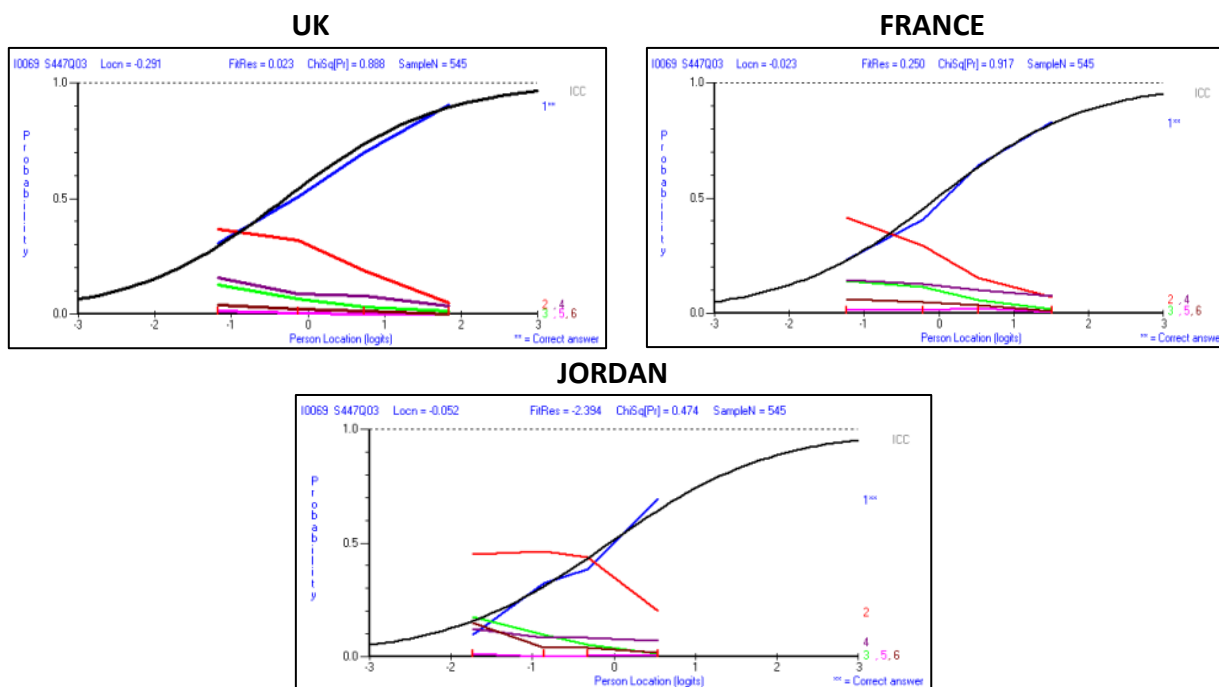
Item 69 was somewhat easier for UK students (item location = -0.32) than it was for the Jordanians (item location = -0.03) or the French (item location = 0.08) as shown in Figure 7.17 despite the non-significance of overall DIF.

Figure 7.17: Item locations per country for item 69



The results of the distractor analysis are displayed below in Figure 7.18. For all three countries, the proportion of participants choosing the key, i.e. option 1, increases with person ability while the proportions of selecting any of the other options, the distractors, decrease with person ability. For all countries, distractor 2 seems to be highly misleading for low ability students. However, the proportion of selecting it decreases in the UK to become nil for high ability students. This distractor keeps a 10% chance of being selected for high ability students in France and greater than 20% of being selected by high ability students in Jordan. Given that this distractor is misleading in all languages, but to different degrees, it is less likely that a translation-related feature is what is attracting students to select it. It is possibly a misconception that is more or less strongly present in a country. A qualitative analysis can provide more evidence for or against this hypothesis.

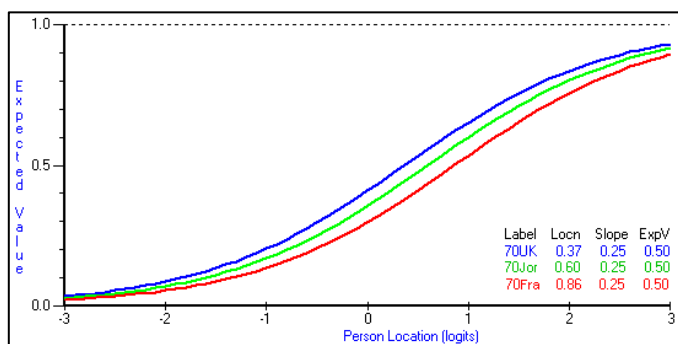
Figure 7.18: Distractor analysis of item 69



Item 70 – S447Q04 (Sunscreens)

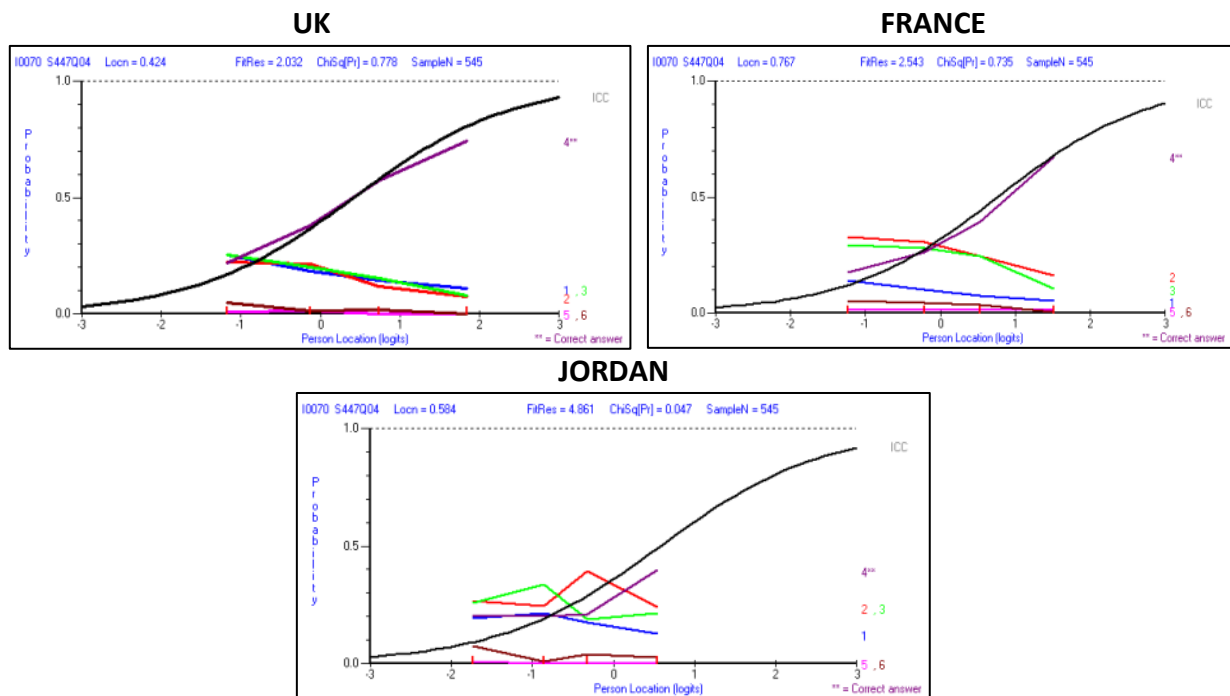
Although not flagged for DIF, item 70 was somewhat easier for the UK students (item location = 0.37) than it was for the Jordanians (item location = 0.60) or French (item location = 0.86) as shown in Figure 7.19.

Figure 7.19: Item locations per country for item 70



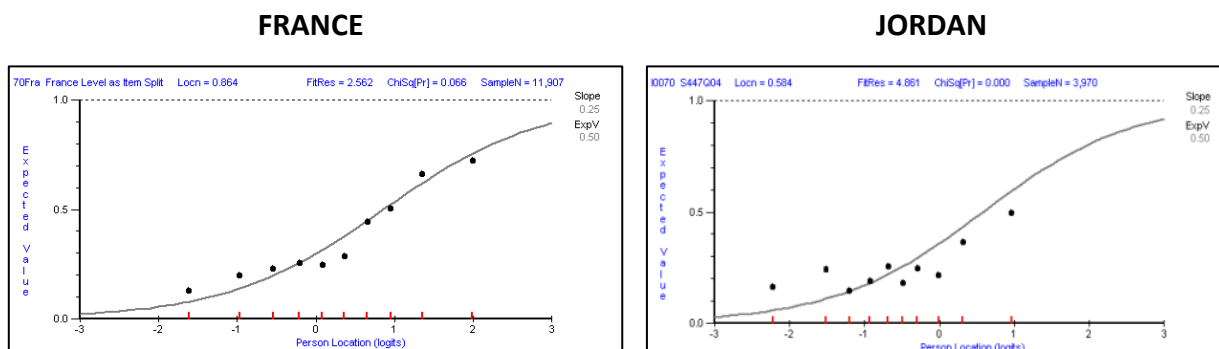
The results of the distractor analysis are displayed below in Figure 7.20.

Figure 7.20: Distractor analysis of item 70



For the UK and France, the proportion of selecting the key, option 4, increases with person ability while the proportions of choosing any of the other options, the distractors, decrease with person ability. The UK DPC fits the model better than the DPC of the other countries do although options 1 and 3 never approach 0 for high ability students. The France DPC reveals that a relatively high proportion of students selected distractors 2 and 3 including students with moderate abilities with the proportion of selecting distractor 2 remaining higher than 20% for high ability students. This suggests that random guessing may have been involved with distractors 2 and 3 in France. The DPC of Jordan shows a strong misfit to the model; a very low discrimination pattern and a great possibility of random guessing. These hypotheses are supported by the ICCs of this item in each country presented in Figure 7.21 below. For both countries, the country ICCs suggest that low ability students performed better than expected by the model while high ability students performed worse.

Figure 7.21: ICC curve of item 70 in France (left) and Jordan (right)

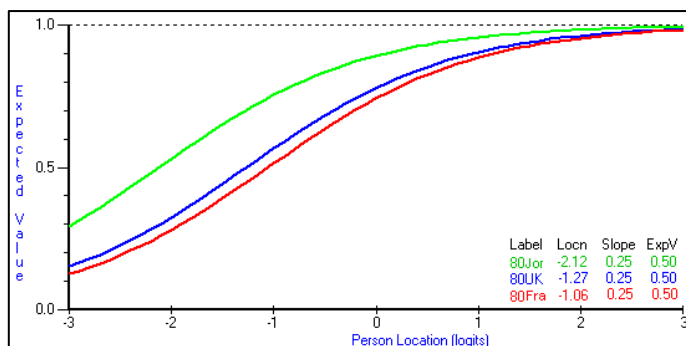


A qualitative analysis would help explain whether there is a differential impact of translation on item difficulty or whether the source leading to different DPCs is not language-related.

Item 80 – SQ47702 (Mary Montagu)

Item 80 was somewhat easier for Jordanian examinees (item location = -2.12) than it was for the UK (item location = -1.27) or France (item location = -1.06) as shown in Figure 7.22.

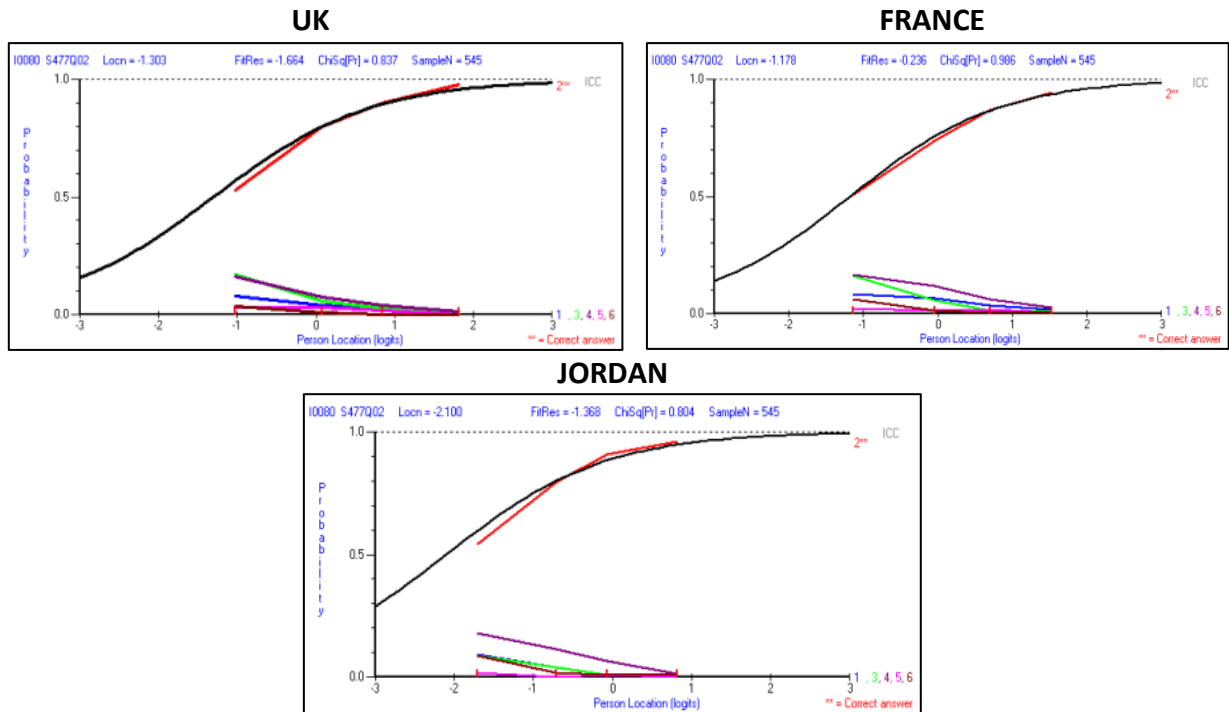
Figure 7.22: Item locations per country for item 80



The results of the distractor analysis are displayed below in Figure 7.23. For all three countries, the DPCs show a good fit to the model. The proportion of students selecting the key, option 2, is higher and increases with person ability while the proportions of all the other options, the distractors, are lower and decrease with person ability. For this item, the distractor analysis does not reveal any differential functioning of the distractors in the different

countries. Indeed, the item is one of the easiest items (fourth easiest amongst the 103 items) in the test and does not seem to have been problematic in any way.

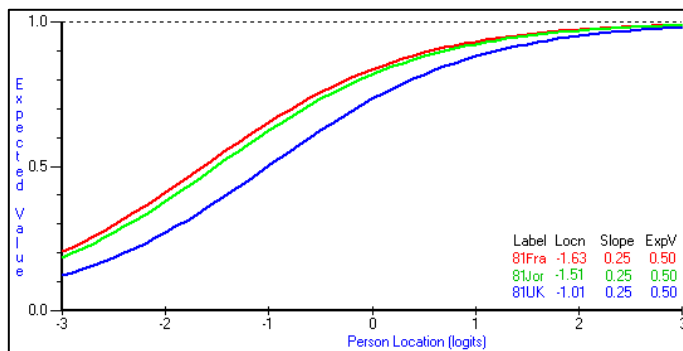
Figure 7.23: Distractor analysis of item 80



Item 81 – SQ47703 (Mary Montagu)

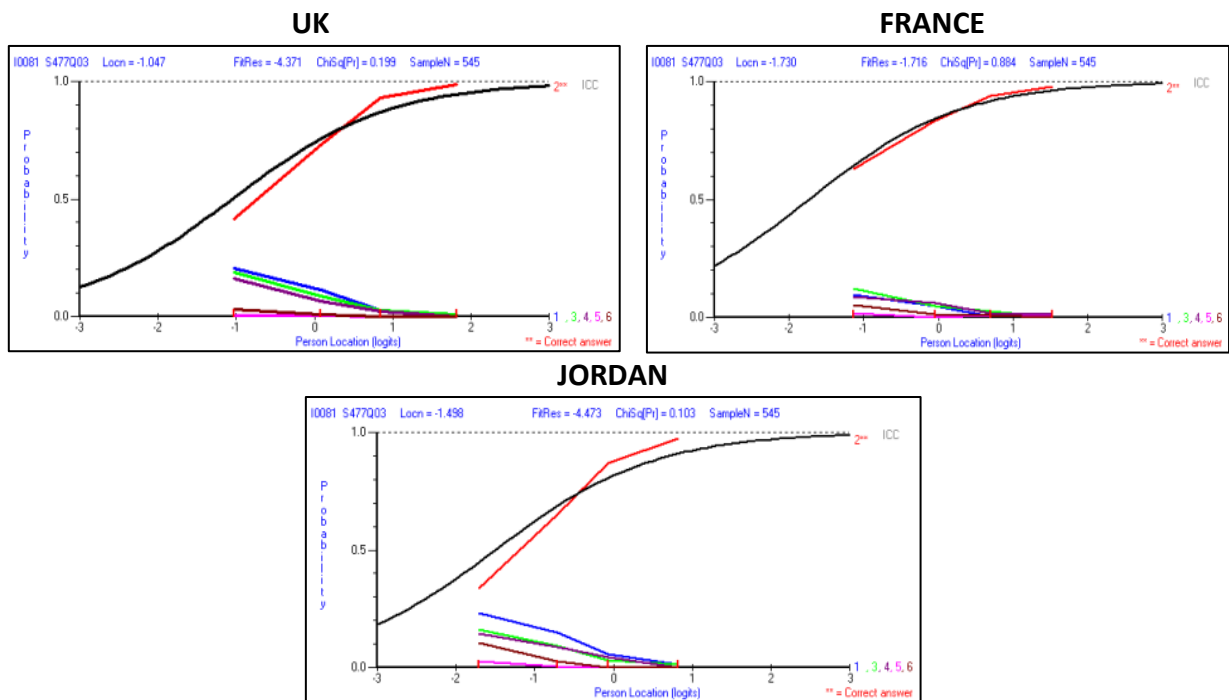
Item 81 was somewhat more difficult for the UK sample (item location = -1.01) than it was for the Jordanians (item location = -1.51) or French (item location = -1.63) as shown in Figure 7.24.

Figure 7.24: Item locations per country for item 81



The results of the distractor analysis are displayed below in Figure 7.25. For all countries, the DPCs show a good fit to the model. The proportion of individuals selecting the key, option 2, is higher and increases with person ability while the proportions of choosing any of the other options, the distractors, are lower and decrease with person ability. Similarly to item 80, the distractor analysis does not reveal any differential functioning of the distractors in the different countries. The item is also one of the easiest in the test (seventh easiest).

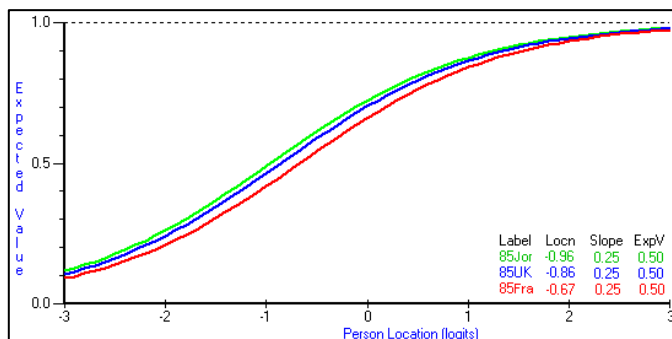
Figure 7.25: Distractor analysis of item 81



Item 85 – SQ485Q03 (Acid Rain)

Item 85 seems to have been somewhat easier for the Jordanians (item location = -0.96) than it was for the French (item location = -0.67) or the UK (item location = -0.86) as shown in Figure 7.26.

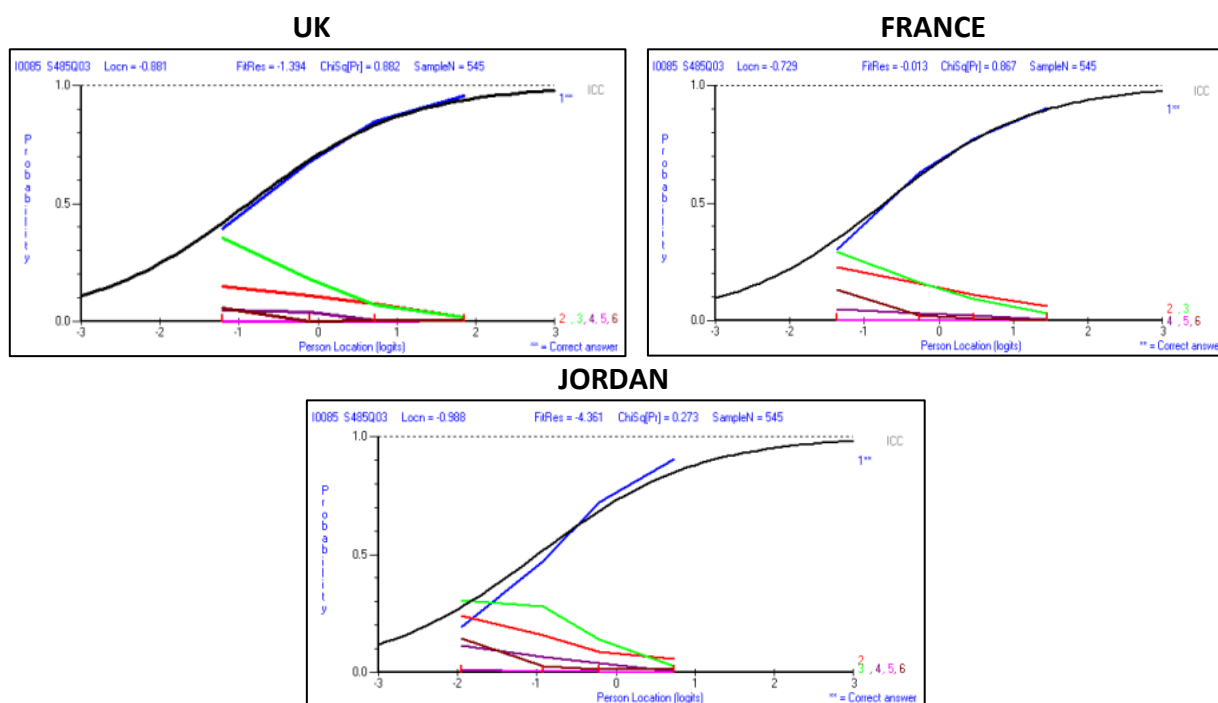
Figure 7.26: Item locations per country for item 85



The results of the distractor analysis are displayed below in Figure 7.27. For all three countries, the proportion of the key, option 1, is higher and increases with person ability while

the proportions of all the other options, the distractors, are lower and decrease with person ability. The proportion of choosing distractor 3 is relatively higher than that of the other distractors especially in Jordan. However, on the whole, the DPCs do not point to any differential distractor functioning across countries.

Figure 7.27: Distractor analysis of item 85



Distractor analyses allow the gathering of more evidence on the way items function in different groups by observing the behaviour of the key and the distractors across the groups. This additional analysis was not only carried out on DIF items. Multiple choice items that may have not been flagged for DIF might still show different behaviour of distractors and key across countries. Some distractors may be more confusing in some countries because of the language difference or other factors (curriculum, familiarity with the context, misconceptions, etc.), yet on the whole, the item might have worked equally well in all countries. Given the emphasis of this research on the impact of language on difficulty at item level, this analysis

allowed access to evidence at a deeper level. Ten items were analysed using distractor analysis irrespective of whether they have been flagged for DIF. Evidence for differential distractor behaviour was recorded in all DIF items and in four out of seven non-DIF items. A thorough qualitative analysis of the items can reveal the sources of differential distractor functioning provided the methodology adopted is reliable and can detect the source of the variation in the distractor behaviour such as subtleties in languages. This point is elaborated further in the next chapter.

CONCLUSION

The emergence of DIF in an item is often a sign of multidimensionality in that item (Marais & Andrich, 2008b). In this chapter, a DIF analysis was carried out using a two-way ANOVA of response residuals (Hagquist & Andrich's, 2004) to investigate whether the unidimensionality suggested by the PCA in Chapter 6 (Section 6.7.4) was maintained at item level. This approach has not been used earlier to examine bias across language versions of the same test. After eliminating artificial DIF produced as a 'by product' of the procedure to detect it (Andrich & Hagquist, 2012, 2015), twenty-three items were flagged for real DIF. This finding was consistent with previous studies investigating DIF in PISA (e.g. Grisay & Monseur, 2007; Grisay et al., 2007; Huang et al., 2014; Kreiner & Christensen, 2014; Le, 2009; Oliveri & Ercikan, 2011; Xie & Wilson, 2008).

When comparing the relative number of DIF items advantaging or disadvantaging countries, France seems to be the most disadvantaged, having the highest number of items with bias against it and lowest number of items advantaging it. The opposite is true for Jordan which seemed to be overall benefiting from a positive bias towards it. The average difficulty of DIF

items per country conveyed a similar picture, France having the highest mean item difficulty and Jordan having the lowest. In addition, the mean difficulty of DIF items of the UK was closer to that of Jordan. However, a scatter plot of the location of DIF items in pairs of countries suggested that UK and France are closer than UK and Jordan or France and Jordan. This latter finding is consistent with findings of Grisay and Monseur (2007). The amount of DIF between versions of ‘cousin languages’ (i.e. English and French) and between countries with similar GDP is less than that between versions of distant languages (English-Arabic or French-Arabic) and between countries with different GDP (UK-Jordan or France-Jordan).

Despite evidence of DIF at item level, DIF did not seem to have a significant impact on the test as a whole. This corroborated the findings of Oliveri and Ercikan (2011). The approach adopted combined evidence from two studies. The first was a repeated-measures ANOVA comparing mean item locations of DIF items across countries. The second study consisted of a comparison of mean person estimates before and after splitting items in each country. This approach has not been used in any of the previous studies investigating DIF in international assessment. The first study pointed to insignificant differences between means of DIF items at test level while the second indicated that deviations of person estimates before and after splitting items were significant yet very small. DIF in items balanced out and no country was significantly advantaged or disadvantaged at test level.

Indeed, this could be a technique adopted to get around inevitable DIF at item level. Hence, if DIF was a manifestation of bias in some items, it seems to have been neutralised statistically at the test level. If it was a manifestation of multidimensionality with the additional dimension(s) advantaging or disadvantaging a particular country then only a thorough

qualitative analysis would be able to uncover the additional factors contributing to the variance.

Distractor analyses on multiple choice items were carried out as complementary studies to the DIF analyses to investigate whether distractors behaved differently across countries and whether some countries were more or less attracted by some distractors. These analyses were carried out on released simple multiple choice items irrespective of whether they were flagged for DIF or not. Ten items were analysed, three of which had been flagged for DIF while the seven others had not been. All DIF items and four of the non-DIF item showed evidence of differential distractor functioning. Distractor analyses allowed the gathering of additional evidence on the functioning of items and would be more informative if they are followed up by qualitative analyses of the items.

Stout (2002) pointed out that ‘one of the subtle ways that DIF has been compartmentalized is the almost total disconnect that has evolved between substantive (content-based) and DIF (statistical) approaches to the understanding and practice of test fairness’. DIF analyses are rather simple analyses; however, they present several limitations. First, these analyses only point to inconsistent behaviour of specific items across groups; they do not allow the identification of factor(s) leading to the emergence of DIF. In this context, we hypothesize that language is one of the explanations; however, DIF could be caused by different factors or a combination of them: cultural differences, curriculum, nature of educational systems, school types, etc. Besides, it is important to keep in mind that DIF is not an absolute measure. DIF results from the interaction between some items in relation to the rest of the items. In cases where all items would be equally biased in one direction, for instance UK in this context, DIF would not occur. Distractor analyses were used as complementary analyses. However, neither

of these techniques can answer my research question without a thorough qualitative investigation of the items. The next chapter presents the various techniques available for carrying out these analysis and discusses their limitations.

CHAPTER 8

Judging Test Demands across Languages

‘No amount of painstaking translation will quite eradicate the Englishness of thought which underlies it.’

(Jones, 1988, p.179)

INTRODUCTION

Chapter 7 discussed results of the DIF analysis carried out on data from PISA Science items of the 2006 cycle from three countries: UK, France and Jordan. DIF was detected in twenty-three out of a hundred and three items despite balancing out at test level. The DIF analysis was complemented by a distractor analysis of all ten released simple MCQ items. The distractor analysis provided additional insight into whether the distractors of simple MCQs functioned the same way across countries irrespective of whether the item showed significant DIF or not. Out of the ten released simple MCQ items, seven showed differential distractor functioning, of which only three had been flagged for DIF.

DIF analyses only detect items which show significant differences in the item mean score across countries given the overall test score. Hence, if the entire test is biased in one direction, DIF analyses would not be able to pick this bias up. Besides, neither the DIF nor the distractor analyses could identify the source(s) leading to the differential behaviour of items or distractors across countries. However, these analyses were useful in flagging items worth following up more substantively.

While DIF and distractor analyses are statistical in essence, understanding the sources leading to differential behaviour of items requires a qualitative explanation and is usually determined by expert judgement (Sireci et al., 2005, p.103). Indeed, the difference of the

nature of the analyses reflects the distinction made by Pollitt et al. (2007) between difficulty and demand of an item. Item difficulty is the empirical proportion of examinees answering a given question correctly. DIF analyses are *post hoc* analyses and are based on the significant differences of item difficulty across groups and a similar logic applies to distractor analyses, where the significant differences are flagged at distractor level across groups.

Demands refer to the cognitive processes that examinees experience when solving a given item. Demands are essentially qualitative and hence cannot be investigated using quantitative techniques. They could be of different forms; for instance, the amount of reading or writing required, the load imposed on working memory, the load imposed on long-term memory, etc. (Pollitt et al., 2007).

Early approaches to judging demands of educational assessment items did not rely on robust methodologies and did not yield conclusive results. Several models have been proposed in the literature, the more recent ones being the question answering process (QAP; Pollitt & Ahmed, 1999), the CRAS framework (Pollitt et al., 1998) and the Task Difficulty Framework (Lumley et al., 2012). With the exception of the last one, all major studies comparing judgements of item demands in educational assessments occurred in the British context, to compare demands of exam papers produced by different exam boards or of different subjects or different years.

These models were criticised for various practical and methodological reasons, mainly for having low consistency in the ratings. Besides, with the exception of CRAS, all previous approaches and models were criticised for including aspects that did not relate directly to demands. The last model, Lumley et al.'s (2012) Task Difficulty model, was proposed specifically for predicting the difficulty of PISA reading tasks. It was applied to different

language versions of PISA reading tests and yielded high inter-rater reliability scores. However, no other studies used this framework.

Comparing demands of different language versions of the same test requires a model that examines the language variable in an in-depth way. The model should include a variable that looks extensively at the effect of translation and adaptation. So far, as discussed in Chapters 3 and 5, judgement of the quality of translation and the equivalence of translated versions has been based on expert judgement. However in this chapter, I discuss various issues with this methodology. I review the literature on cognitive load which focuses on the cognitive demands involved in the interaction between an individual and a task. I apply this theory in the context of PISA items and highlight some of the variables that are inherent in the language and that would not be easily detected by the available methods.

This chapter is divided into seven main sections. The first distinguishes between the concepts of difficulty and demand. The second section describes various types of demand and illustrates some of them. The third section summarises various approaches used in comparability studies of demands. The fourth section describes three main models used for judging item or task demands: (1) QAP model, (2) CRAS framework and (3) Lumley's Task Difficulty framework. The fifth section examines the literature on survey development for some additional insight on features making questions more demanding. The sixth section explores cognitive load theory and uses examples of PISA science 2006 released items to illustrate its application in comparing demands of educational assessments across cultures.

8.1 Difficulty versus demands

There is some confusion in the literature between the terms ‘difficulty’ and ‘demand’. In this thesis, I adopt the distinction that Pollitt et al. (2007) clearly described and which I briefly discussed in Chapter 2. Difficulty refers to the empirical proportion of individuals who got a given question correct. A demand designates a cognitive process placed on a given person when carrying out a task. This distinction was well portrayed in an example Pollitt et al. (2007) gave about two TIMSS items administered in England and Scotland (see Figure 8.1).

The questions in Figure 8.1 were identical in England and in Scotland and administered to the same age group. Hence, the cognitive processes required for solving each of these questions were the same across countries; in other words, the demands were the same. However, Pollitt et al. (2007) reported differences in success rates across countries. In Scotland, these two questions were equally difficult; that is, the percentage of Scottish children getting these questions correct was respectively 75% and 76% (TIMSS, 1996b). However, in England the success rates differed on these questions and were respectively 59% and 79%.

These results suggest that English students found the first question more difficult than the second question. Moreover, they found the first question more difficult than their counterparts in the Scottish cohort found it. These results have led to intense criticism of the English educational system as reported in Pollitt et al. (2007); however, Whitburn (1999) related the difference in performance between the countries to the difference in classroom experience up to the test administration date between the two countries. In Scotland, students learn strategies to solve questions similar to question 1 earlier than

students in England making English students unable to perform well on the first question by that age.

Figure 8.1: TIMSS items

1 Subtract: $\begin{array}{r} 6000 \\ -2369 \\ \hline \end{array}$	2 Write a fraction that is larger than $\frac{2}{7}$
A. 4369 B. 3742 C. 3631 D. 3531	Answer: _____

Source: TIMSS, 1996a, pp.105-107

8.2 Types of demands

Nine main examination features could affect the demand of a given question (Pollitt et al., 2007). These are presented below:

- a. Duration of an assessment: The time allotted for an assessment could increase or decrease overall demands depending on the amount of work required during this time. When a task is given just enough time to be completed, the overall demand is relatively higher compared to the same task given more time than it requires.
- b. The amount of work to be carried out: Typically, overall demand increases with the amount of work required on a given task. Tests with fewer questions in a given time are normally less demanding than tests with more questions to be solved in the same time. However, one should keep in mind that this is not as straightforward as this relationship depends on the nature of the questions. For instance, examiners may choose to break down a demanding question into several simpler lead questions. The example below shows how a single demanding question requiring several steps can be divided into simpler less demanding questions.

Figure 8.2: Relative amount of work per question

A single demanding question	Three simpler questions
Sarah bought 5 red pens and 4 blue pens. She gave Pete 2 pens of each colour. How many pens was Sarah left with in total?	Sarah bought 5 red pens and 4 blue pens. She gave Pete 2 pens of each colour. a) How many pens did Sarah buy in total? b) How many pens did Sarah give to Pete in total? c) Deduce how many pens Sarah was left with in total?

- c. The amount of reading or writing required in a given time: Typically, overall demand increases with reading and writing load. It is however worth noting that longer questions may sometimes decrease overall demand when they provide additional instructions or clues to solving the question.
- d. The level of reading difficulty: The higher the reading difficulty, the higher the cognitive demands required to process the material being read.
- e. The amount of question choice: This is relevant to examinations providing students with a number of questions from which students are required to answer a few. Overall demand is hard to predict in this case as it is influenced by various interacting factors: students' expectations, preparation and curriculum coverage.
- f. Demand made on long-term memory: Different subjects impose different levels of demand on long-term memory; so do test formats and question types. For instance, open book questions decrease the level of demand made on long-term memory.
- g. Demands on working memory: Complex questions such as maths problems requiring various steps impose higher demands on working memory.

- h. The nature of the cognitive processes required lead to different levels of demands. Following Bloom's (1956) taxonomy, questions targeting the knowledge level, such as labelling the digestive system, impose fewer demands than questions tackling the analysis level on the same subject, and requiring students to analyse data from a table and explain digestion.
- i. Stress: Working memory is highly affected by stress experienced during examinations. Spielberger (1972) distinguished between affective stress associated with test anxiety and results. This type of stress tends to vary amongst people depending on their character – this is *trait anxiety*. The other type of stress is cognitive in nature and is caused by high demands in a given context, such as a particularly complex question in an exam – this is referred to as *state anxiety*.

The list of types of demands described above is not exhaustive. One can, for instance, also add test-wiseness arising from familiarity of examinees with given types of items (Bunting & Mooney, 2001). The demands mentioned above contribute to increasing or decreasing overall demand experienced by examinees during a given assessment. In the case of comparing the demands of language versions of the same examination, such as comparing the demands imposed by different versions of the same PISA test in different countries, most of these demands are the same across languages (such as the duration of the assessment, the number of options, the format of the questions, etc.).

Of particular interest in this type of comparison are the types of demands described in (e) 'the amount of reading and writing' and (d) 'the reading difficulty'. In international assessments, language versions of the same test could impose different reading and writing demands. For instance, Eivers (2010) found that the Finnish versions of PISA tests were on average 8% longer than the English versions, the Irish versions were 11% longer and the German versions were 17% longer. This suggests that the Finnish, the Irish and the German versions are likely to have imposed additional demands than did the English version of the same PISA test. Similarly, Huang et al. (2014) found that the English

version of PISA science 2006 test was significantly longer than the Chinese version and reported higher language demands in the USA compared to mainland China. Ibrahim (2009) argued that PISA reading tests are more challenging in Arabic because of the language distinctive features.

Comparing tests on a particular aspect, such as length of a test looks straightforward; however, judging demands especially when it comes to the cognitive processes involved is much more complex. Often, various aspects of demands interact and compensate for each other in unpredictable ways. Initial explorations of question demands were carried out informally while more recent ones followed more systematic procedures. Despite the improved design of these studies, they still failed to judge demands reliably. The next section describes various approaches developed to investigate demands placed on examinees when taking exams.

Cognitive processes involved in solving a particular task vary with the nature of the task and the subject and are influenced by its length, its complexity and examinees' experience of the particular problem type or concept. The following section describes the main types of demands imposed during assessment, as outlined by Pollitt et al. (2007).

8.3 Approaches for judging question demands

This section describes bottom-up approaches applied to explore demands of question papers. Early studies were conducted in Britain to compare demands of question papers across examination boards as well as in an aim to maintain standards across subjects and across years. The approaches lacked sophistication prior to the introduction of the National Curriculum in 1992. Their design improved thereafter and became more systematic. This section summarises these studies since the first informal investigations of question

demands were carried out. Later, more defined designs are discussed and then the application of the construct elicitation technique in the field of educational assessment is presented. The last part of this section discusses systematic methods used in a few comparability studies in the PISA context.

8.3.1 Informal investigations of question demands

Before 1992, only a few studies were published comparing demands imposed on students during examinations. Most of these studies were carried out in the British setting comparing high-stakes examinations across examination boards. These studies used informal methods to identify specific demands. Some of the methods used included the use of questionnaires to gauge views on the relative demand of English language O-level papers (Massey, 1979), the use of participants' ratings of demands of A level economics papers (Houston, 1981) and the analysis of assessors' comments and analyses made prior to scrutiny sessions (Evans & Pierce, 1982).

Pollitt et al. (1985) categorised features making Scottish Ordinary Grade examinations questions more difficult into three categories that would be today described as aspects of demands: (1) subject or concept difficulty which reflects the inherent demands imposed by the content, (2) process difficulty which is associated with the cognitive demands required to achieve the task, (3) question difficulty which relates to the way the task is formulated and presented. McLone and Patrick (1990) adopted a more complex method to compare the demands of two mathematics examinations and used a matrix based on Griffiths & McLone (1979). Judges were asked to rate examination papers on a scale of 0 to 3, based on specific statements they were given, for example, *How far does the question define in detail the procedure which the candidate should adopt?* The report highlighted the

challenges associated with having a shared interpretation of the statements amongst judges, the poor consistency of judgements and the application of scales.

8.3.2 Systematic investigations of question demands

More systematic approaches to comparing demands were adopted in studies carried out in light of the introduction of the National Curriculum (age 5 to 16) in the UK. These studies were carried out on GCSE maths, English and science (Jones, 1993), history (Stobart et al, 1994), and geography (Ratcliffe, 1994). Judges were familiarised with syllabus material. They were then trained to make judgements by practising the use of defined dimensions to rate the demands of syllabuses (Jones, 1993) or by rating syllabus demands against a number of factors based upon previous research by Pollitt et al. (1985), Stobart et al. (1994) and Ratcliffe (1994). They were then asked to rate exam papers using specific components of demands and sometimes relative to overall demand. Some of these studies did not lead to any clear conclusions about the comparability of demands (Jones, 1993). Pollitt et al. (2007) held that although researchers and judges often reported being satisfied with the methodology, this did not make the methodology valid as often satisfaction is a reflection of how well and easy the judges were able to accomplish the task. Some studies conducted on modular and non-modular syllabuses in A level²² biology used methods that involved judges writing qualitative reports based on a number of factors (e.g. content, processes, question or stimulus difficulty) instead of making quantitative judgements and researchers finalising their statements (D'Arcy, 1997). This method was later criticised for being highly prone to bias arising from the greater familiarity of judges with their own syllabuses and the way researchers summarised judges' comments (Jones, 1997). Attempts

²² Examinations taken at age 18 in England.

to use overall demands to compare exam papers were deemed of little value (Walker et al., 1987).

8.3.3 Construct elicitation technique

A new approach at that time, the construct elicitation technique, became popular in the 1990s for judging demands of exam papers. This method was based on Kelly's (1955) seminal work in psychology. Kelly contended that reality is subjective and that one's perception of the universe determines our own set of constructs. The construct elicitation technique involved a procedure that allowed the practitioner or researcher to understand how a particular individual interpreted the world by bringing out their construct repertoire.

The construct elicitation technique was applied in various educational assessment studies (e.g. Adams & Pinot de Moira, 2000; Fearnley, 2000; Gray, 2000; Greateorex et al., 2002; 2003; Pritchard et al., 2000). It consisted of presenting a judge from each participating exam board with assessment materials (that is, specifications, questions papers and mark schemes) from pairs of syllabuses and then requiring them to record a minimum of three similarities and three differences in the demands imposed on examinees while being assessed. This first step enabled examiners to understand what constituted demands and derive a scale with specific constructs (Gray, 2000).

The definitive set of constructs was then agreed upon jointly amongst participants and finalised in the form of questions or statements; each with a title and rating anchors. The final set of constructs was compiled in the form of a questionnaire with an additional item requesting the rating of overall demand of the exam paper. Edwards and Adams (2003) allowed examiners to modify their original ratings following the cross-moderation session during which examiners were exposed to student scripts. This last step was criticised by

Pollitt et al. (2007) for conflating the conception of ‘demand’ with that of ‘difficulty’. The former refers to the needed cognitive processes that question-writers decided to include in a question prior to its administration while the latter is an empirical measure that occurs *post hoc*.

According to Pollitt et al. (2007), although construct statements conveyed accurately the constructs elicited from judges, the method presented several problematic features in terms of its effectiveness in evaluating the types and levels of demands included within an exam paper. For instance, Pollitt et al. wondered about the extent to which some of the construct statements referred explicitly to demands or were associated to them in any way. Besides, interactions amongst demands may impact overall demand in different ways. A common finding amongst these studies is that significant differences are observed between ratings of constructs but these have been reported to be as low as only a few (e.g. Fearnley, 2000; Greateorex et al., 2002).

Besides that, none of these studies reported finding substantial differences in the rating of the overall demands although Baird (1999) analysed Fearnley’s (2000) published data using other techniques and reached different results. She found that a small number of constructs explained over half of the total variation in overall demand ratings and that these excluded the constructs identified by Fearnley (2000). Hence, there is a lack of consistency amongst studies looking at the same data in terms of identifying demands of an assessment. Also, these aspects were reported to be inconsistent across subjects; in other words, different subjects tended to have different types of salient demands.

8.3.4 Systematic comparisons of language versions of PISA

Very few studies used systematic qualitative approaches to examine the differential demands of items in PISA. Arffman (2012) examined the quality of translation of the Finnish version of PISA reading 2000 assessments while Huang et al. (2014) investigated the sources of DIF by comparing the English and Chinese versions of PISA science 2006.

Arffman (2012) recognised the need to develop better techniques for carrying out proper reviews of items. She proposed a systematic linguistic comparison method. The method involved two independent reviewers: a linguist with an experience in test translation and a trilingual (Finnish, English and French) translator and linguist with some experience in translation. The method comprised three stages. The first stage consisted of a careful reading of the text and items in the assessment to get familiar with the test. The second stage involved consulting national reports available on the items including DIF items. According to the author, this stage provides indirect information on the cognitive demands of the item. The third stage includes a systematic comparison between the Finnish version and the source versions (English and French).

The comparison between the Finnish and source version of PISA was based on the morphology, the syntax, the semantics and the pragmatics of each of the three language versions. Factors affecting items and text difficulty such as familiarity, explicitness and transparency, were evaluated across the languages. Arffman's attempt is a step forward into examining closely the translation quality of PISA assessments across languages. However, her method has several weaknesses.

First, the results are likely to suffer from low reliability between the reviewers just like the studies carried out in the UK. The author did not give any information about the extent to which the reviewers agreed on the translation problems they identified; nevertheless, she

recognised the importance of having more than one reviewer involved and the challenge of finding suitable reviewers for the task with the right language and subject expertise. Second, the method relied on judgements made by multilingual reviewers who most likely would have different levels of proficiency in the different languages and thus introducing some bias in the judgement (Hambleton, 2005) as outlined in Chapter 3 (Section 3.5.2.a). A third problem with this method is the lack of clarity of how the researcher could identify cognitive demands based on the DIF items. The researcher mentioned making ‘informed guesses’ (p.5); this approach is weak and cannot be transferred to other studies.

Another study carried out by Huang et al. (2014) compared the DIF items of two English and two Chinese versions of PISA science 2006 in terms of language, curriculum coverage and cultural bias. A content analysis methodology was employed using a bilingual reviewer for the analysis of language demands. These were related to difficult vocabulary, passage length, grammatical structure and contextual meaning of the passage. The curriculum coverage and the cultural bias were judged using instruments with problematic items administered to a panel of around twenty reviewers with a balance between English-speaking and Chinese-speaking reviewers. The reviewers included ten teachers and twenty students. The input of teachers was considered valuable as they represented subject experts. Students were considered in this study so as to gain insight into their perceptions of what constitutes a demand.

The study did not provide information about the reliability of judgements made by teachers and students. Hambleton (2005) pointed to issues with using monolingual samples to judge the equivalence of tests (see Chapter 3, Section 3.5.2.d). In Huang et al.’s (2014) study, American and Chinese teachers and students belong to very different cultures separated by a large number of variables such as curriculum, expectations, socio-economic and political

factors. These introduce bias in the ratings which makes them likely to suffer from low reliability just like the studies previously described.

Early approaches to judging demands have either used weak methods or have relied on bottom-up approaches which often led to making judgements on aspects that are not necessarily demand components. Later studies carried out in the PISA context suffered similar problems. The field is in need of a framework of what is qualified as demands in the question answering process. The next section describes three main models for judging demands.

8.4 Models for judging demands

Various models for judging the demands of items and tasks have been suggested in the literature (see review by Dhillon, 2003). Some of these models focused mainly on the cognitive demands induced by questions and some of them attempted to forge a predictive relationship between challenging aspects of a task or an item and its difficulty. This section presents three models attempting to judge demands of items in the chronological order of their development as this somehow portrays the evolution of thinking in the field: (1) Pollitt and Ahmed's Question Answering Process, (2) CRAS, (3) Lumley's task difficulty framework.

8.4.1 Pollitt and Ahmed's Model of the Question Answering Process (QAP)

Sternberg (1977) proposed a model depicting a sequence of six 'cognitive components' involved in response generation when a candidate is faced with non-verbal reasoning questions: (1) encoding, (2) inference, (3) mapping, (4) application, (5) justification, and

(6) response. Sternberg used non-verbal tasks to eliminate the language variable when examining the cognitive processes involved in solving a given question.

In the late nineties, Pollitt and Ahmed (1999) extended Sternberg's (1979) model to verbal tasks. They proposed a model, the Question Answering Process (QAP), which drew on various fields of cognitive psychology, such as learning, memory, problem solving, attention and language comprehension. The model consisted of 6 stages or components describing the sequence of conscious and unconscious mental stages involved in producing a response. When learning a new concept (stage 0 in the question answering process model), students reach a certain understanding by constructing mental models based on prior experiences and knowledge. In examinations, students (1) read and understand the question, (2) search their mental models for relevant parts, (3) interpret the mental models to match them with the question; (4) find a solution to finally (5) write an answer.

Pollitt and Ahmed identified various cognitive domains involved in each stage of the process. These are respectively learning, language comprehension, memory, attention, problem-solving and language production. Whilst candidates may have mastered the concept being assessed, they may still face difficulties at various stages and these may vary in nature depending on the cognitive domains activated at each stage (Pollitt & Ahmed, 1999, 2000). In fact, students may be hindered by different features embedded in the questions asked.

The fundamental role of language was clearly described in the QAP model as it is directly involved in at least two stages: the first when examinees read the question and the last when examinees write an answer. Since the focus of this research is more on the questions rather than student responses, I will be discussing the challenges examinees face in the first stage of QAP; that is, when examinees are reading the question.

Pollitt and Ahmed (1999; 2000) argued that problematic features in items are most influential at this stage. Students often fail to answer a question properly because they have misconstrued what the task was about rather than misunderstood the construct being evaluated. As discussed previously in Chapter 2, different languages may present students with different types and levels of challenges because of their idiosyncrasies. This makes them less comparable in terms of demands.

A large body of research examined the impact of readability on students' performance in examinations (Kirsch et al., 1998; Rupp et al., 2001; Oakland & Lane, 2004). The following paragraph draws on research carried out in cognitive psychology and explains briefly how understanding written material occurs at the cognitive level.

8.4.1.a Making Meaning While Reading

The cognitive processes involved in reading were described by various cognitive psychologists. Cognitive theory suggests that students comprehend the language they read in texts by actively building coherent mental representations (Garnham, 1987) which include pre-fabricated frameworks or schemas depicting specific contexts or events. When a reader is faced with an unseen text, specific pre-constructed schemas are activated allowing the building of mental representations of what they read to make sense of it. Pollitt and Ahmed (2000) maintained that if one considers assessment questions just like any other textual material, similar processes are expected to apply. Readers attempt to make sense of the question by building mental representations of the task to be achieved.

Nevertheless, Pollitt and Ahmed recognised that reading a question in an examination setting entails at least two variables that are not involved when reading an unseen text in a 'normal' setting: (1) examinees' expectations while reading and (2) reading in an assessment setting. These variables are described below.

When reading an unseen text, the reader has no specific expectations as the text could be about an infinite number of topics. Hence, the mental representations generated are based on ones formerly built from experience or prior knowledge but these have not been envisaged by the student prior to reading. They were simply activated during the reading process. However, when reading an examination question, the candidate has certain expectations prior to the exam about the content to which this question is related. This content is defined by the curriculum and thus, new mental representations emerge but these inevitably take into account the ones previously constructed during the learning process.

In addition, reading a question cannot be dissociated from the setting in which this activity is taking place – the examination setting (Pollitt & Ahmed, 1999). Thus, the mental representations emerging while a candidate is reading a question inevitably incorporate various examination-related schemas, such as the expectations of the candidate, the candidate's perceptions of examiners' expectations, prescriptions for action, the mark allocation, familiarity with the examination, examination technique, etc. These are not activated when a student is reading an unseen text and seem to bias reading a question by making some parts more salient than others (Pollitt & Ahmed, 1999; 2000).

Despite the presence of these two distinguishing factors, Pollitt and Ahmed (2000) hold that it is reasonable to suppose that reading an unseen text and reading a question activate comparable mental processes as both activities involve making meaning of a message encapsulated in words and syntax. This means that features challenging the understanding of a question are features hampering the formation of adequate mental representations. Hence, identifying these features is essential when judging specific and overall demands placed on examinees by a given question. Pollitt et al. (1985) sought features making

Scottish examinations harder to understand and called them sources of difficulty (SODs). Similar research was carried out in the rest of UK to identify SODs and sources of easiness (SOEs) in British high-stakes examinations. The next section discusses these studies and describes the most relevant SOEs and SODs to this research.

8.4.1.b Sources of easiness (SOEs) and sources of difficulty (SODs)

Questions could entail different levels of language demands depending on their formulation. This was supported by research reported in Pollitt et al. (1985) which investigated features impeding understanding of Scottish Ordinary Grade examinations. Further research was conducted in the course of the Question Difficulty Project (QDP) by Fisher-Hoch et al. (1997) on English high-stakes examinations (GCSE, IGCSE and O-levels) in five different subjects (Mathematics, Geography, Science, French, English). These studies revealed the presence of various SOEs and SODs in question items. Some of these are construct-irrelevant requiring students' minds to process additional information that is not part of the construct measured (for instance, embedding science items in confusing contexts (Pollitt & Ahmed, 1999)). Others are intrinsic to the construct such as technical terms in science assessments (see discussion in Chapter 2, Section 2.3.1.b).

Fisher-Hoch et al.'s (1997) study focused on SODs essentially. The main SODs identified in the mathematics trials were related to the number of steps required to answer a question, arithmetic errors, recall strategies and recall knowledge, presence or absence of context, spatial representation, and stated principle. Moreover, some of the SODs were common to different curriculum subjects. For instance, command verbs and paper layout were identified as SODs in both mathematics and geography whereas technical terms and everyday language appeared to challenge students in both mathematics and science. Technical terms refer to subject specific terms such as 'perpendicular' and 'hypotenuse' in

maths or ‘osmosis’ and ‘photosynthesis’ in science. Everyday language refers to terms that have a different meaning in everyday context than in a maths or science context. For instance, the term ‘force’ means strength in an informal context while it refers to the capacity to do work or cause physical change in a science classroom context.

Research conducted by Ahmed & Pollitt (1999) suggested that some SODs tended to be more related to the way students tackled different questions. For instance, the sequence of questions, the combination of topics, the display of the mark allocation and the presence or absence of response prompts seemed to make questions more or less accessible for candidates to answer. Other research has focused on different SODs such as the effect of context in science questions (Pollitt & Ahmed, 2000), students’ expectations (Astolfi, 1997; Pollitt & Ahmed, 1999) and students’ cultural background (Abedi, 2004; Solano-Flores & Trumbull, 2003). These factors appeared to mislead students by distracting them from the core questions, confusing them or insinuating some erroneous examiner expectations.

Given the focus of this research on comparing language versions of the same test, the following focuses mainly on language-based SODs which primarily challenge students in the first stage of the QAP model; that is, the reading phase (Pollitt & Ahmed, 1999). Since the cognitive processes implicated in reading questions are very similar to the ones involved in making meaning when reading a question, then similar features obstructing understanding of a text can impede the comprehension of questions. Such features have been thoroughly described in the case of reading science texts and have been detailed in Chapter 2 when describing the language of science.

The language of science seems to present readers with challenges due to its distinctive characteristics (Halliday & Martin, 1993; Chavkin, 1997; Fang, 2006; Ginsguger-Vogel &

Astolfi, 1987; Groves, 1995; Merzyn, 1987; Sutton, 1998). The technical and non-technical vocabulary as well as the complex grammar structure makes science texts difficult to read. Since the acquisition of the scientific language is one of the main instructional objectives of science literacy (DeBoer, 2000; Norris & Phillips, 2003), and I have argued in Chapter 2 that language is part of the construct in science assessments irrespective of the examiner's intention, the characteristic features of the scientific language are likely to appear in science questions. These are likely to also impede the construal of meaning of questions in science assessment and hence increase the reading demands.

The QAP model broke down the process of generating answers into stages and analysed the potential cognitive demands associated with each stage. The model aimed mainly to manipulate the demands imposed by questions. Investigating SODs and SOEs was useful to the extent of understanding what makes questions more demanding for students. However, this approach failed to systematically alleviate construct-irrelevant demands of questions by manipulating SODs. Fisher-Hoch et al. (1997) maintained that attempting to improve the quality of questions by modifying SODs did not necessarily lead to easier questions. Some of the manipulated questions became more difficult.

This finding can be better understood when being juxtaposed with findings which emerged from the literature on judging demands. Demands often interact in unexpected ways (Jones, 1993; McLone & Patrick, 1990) making it very hard to predict difficulty. Besides, SODs are not descriptions of demands but aspects of assessment that make questions more or less difficult. In other words, they are features in items (e.g. double negatives, complex syntax, use of unfamiliar contexts) which have been empirically associated with fewer

students getting particular questions right. They might result in greater linguistic or cognitive demands, but they are not themselves descriptions of demands.

Pollitt and Ahmed recognised the flaw in the use of SODs and SOEs and have developed along with other colleagues the CRAS framework which focused primarily on cognitive demands placed on examinees when answering a question. The next section describes this CRAS framework.

8.4.2 The CRAS framework

Earlier approaches to judging cognitive demands considered ‘demands’ in a broad sense; that is, demands referring to aspects which challenged students during assessments. Pollitt et al. (1998) proposed a framework that enables the evaluation of demands by raters on various features reflected in the name of the framework: the Complexity, Resources, Abstractness, task Strategy and response Strategy (CRAS) framework. This section highlights the development of the CRAS framework, its various scales, its implementation on British examinations and some of its limitations.

8.4.2.a Development of CRAS

The CRAS framework was based on an analytic scale developed by Edwards and Dall’Alba (1981), ‘Scale of Cognitive Demand’, which quantified cognitive demands imposed on students in secondary science classes in Australia and not necessarily in science assessments. Edwards and Dall’Alba (1981) drew on a range of learning and thinking theories (e.g. Ausubel et al., 1978; Bloom, 1956) and the work of Piaget as interpreted by Novak (1977). Edwards and Dall’Alba’s (1981) scale was not widely used outside Australia because it was not designed for judging demands in assessments. Besides, it was very specific to secondary science and was not easily adapted to other

subjects and school levels. CRAS (Pollitt et al., 1998) was devised to be more generic and more prone to be used for judging demands of exam questions irrespective of subject and school levels (Pollitt et al., 1998; Hughes et al., 1998). Earlier research carried out by Pollitt et al. (1985) and QAP (Pollitt & Ahmed, 1999; 2000) was incorporated into CRAS. The framework was refined following feedback of its implementation in judging demands of various tests in different subjects.

8.4.2.b The CRAS scales of demands

The final instrument, referred to as CRAS scales or CRAS framework, consisted of four scales: complexity, resources, abstractness and strategy as shown in Table 8.1 below from Pollitt et al. (2007, p.186). As shown in the table, the 'strategy scale' is split into two scales: the task strategy scale and the response strategy scale. Empirical evidence suggested that different demands underlie the process of coming up with strategies to solve a given task and the process of clearly formulating an answer. In addition, only levels 2 and 4 were described in the framework as opposed to defining them all or defining the extremes only. This was a deliberate decision made by Pollitt et al. (1998) as evidence suggested this format yielded better results.

Table 8.1: The CRAS scales of demands

	1	2	3	4	5
Complexity The number of components or operations or ideas and the links between them.		Mostly single ideas and simple steps. Little comprehension except that required for natural language. Few links between operations.		Synthesis or evaluation is required. Need for technical comprehension. Makes links between cognitive operations.	
Resources The use of data and information.		More or less all and only the data/information needed is given.		Student must generate or select the necessary data/information	
Abstractness The extent to which the student deals with ideas rather than concrete objects or phenomena.		Mostly deals with concrete objects.		Mostly abstract	
Task strategy The extent to which the student devises (or selects) and maintains a strategy for tackling the question.		Strategy is given. Little need to monitor strategy, little selection of information required.		Students need to devise their own strategy. Students must monitor the application of their strategy.	
Response strategy The extent to which students have to organise their own response.		Organisation of response hardly required.		Must select answer content from a large pool of possibilities, must organise how to communicate response.	

Source: Pollitt et al., 2007, p.186

The framework has not been piloted in studies comparing language versions of the same assessment. While this model might be more generic and applicable to a wider range of tasks and subjects, using rating runs the risk of low inter-rater reliability.

8.4.3 Task difficulty

The most recent model for predicting difficulty based on aspects that influence overall demand is that proposed by Lumley et al. (2012) predicting task difficulty in PISA reading tests based on Kirsch and colleagues' work (Kirsch, 2001; Kirsch, et al., 2002; Kirsch & Mosenthal, 1990). Their framework included ten variables potentially impacting overall demand imposed by a task (See Table below, Lumley et al, 2012, p.2).

Table 8.2: Revised PISA reading item difficulty scheme

1. **Number of features and conditions.** This variable relates to the number of features that need to be located in the text in order to provide an adequate response to gain credit for the question.
2. **Proximity of pieces of required information.** This refers to the identification of pieces that need to be put together in order to answer the question. The rating is determined by the proximity of the relevant pieces of information to each other.
3. **Competing information.** This refers to information in the stimulus and/or in the distractors (if multiple choice) that the reader may mistakenly select, or that the reader may generate, because of its similarity in one or more respects to the target information.
4. **Prominence of necessary textual information:** the prominence of the necessary information - that is, information in the text that is needed to answer the question (even if it is not sufficient by itself to answer the question).
5. **Relationship between task and required information:** the relationship between the question (the whole task, including the multiple-choice options where relevant) and the required information - that is, the kind of answer required to gain credit.
6. **Semantic match between task and text:** the degree to which there is a semantic match between the wording of the task and the necessary information - that is, information in the text that is needed to answer the question (even if it is not sufficient by itself to answer the question).
7. **Concreteness of information:** the kind of information that readers must identify to complete a question (Mosenthal, 1998).
8. **Familiarity of information needed to answer the question.** This variable distinguishes tasks that focus on information inside or outside the text, or the text structure, that is close to the experience and concerns of the reader, from those focusing on what is likely to be remote and unfamiliar.
9. **Register of the text.** This refers to the stimulus or part of the stimulus that the reader needs to refer to in order to complete the task. It takes into account the implied relationship between the reader and the text, and the lexico-grammatical density.
10. **Extent to which information from outside the text is required to answer the question.** This variable deals with the extent to which the reader needs to draw on world knowledge, experience or personal beliefs and ideas and opinions in order to answer the question.

Source: Lumley et al, 2012, p.2

Various correlations and regression analyses pointed to three variables being highly associated with the level of item difficulty (8. familiarity of information needed to answer

the question; 5. relationship between the task and required information, 3. competing information in a given task or text). Additional multiple regression analyses pointed to two additional variables contributing to the explanation of variance (10. relevance of information outside the text, 7. concreteness of information).

The purpose of the authors was not to develop a model for judging demands placed on students in PISA reading tests. It only aimed to identify possible sources explaining item difficulty in PISA reading assessments. It is thus not surprising to find that these features did not necessarily refer to demands as defined by Pollitt et al. (2007), that is, they did not explicitly refer to cognitive processes involved in a task.

Lumley et al. (2012) carried out analyses with trained and non-trained raters and reached a reasonable level of agreement amongst trained raters. Based on their findings, the authors proposed increasing the efficiency of the rating scheme by reducing the number of variables to the five variables that resulted in the highest correlations between raters. No other studies adopting this framework have been published yet. The high level of agreement amongst trained raters in this study is encouraging. Besides, the model was used by a bilingual rater (English and Hungarian) to rate the Hungarian version of the test. The study suggested that the Hungarian rater did not have a problem applying the model to the Hungarian version. This finding is also encouraging as the model seems to be suitable to rate reading items in other language. Nevertheless, more evidence should be collected in order to have more solid conclusions about the usefulness of this framework in comparing language versions of the same test.

It is worth noting however that reasonable agreement amongst raters does not necessarily make this model suitable for international comparisons. Of particular relevance to the comparability of language versions of the same test are variables 9 and 6 in decreasing

order of importance. Variable 9 ‘Register of the text’ reduces the linguistic features to a variable of ‘lexico-linguistic density’ which overlooks aspects like word frequency and sentence length discussed in Chapter 2 which vary across languages. It also limits the investigation to the stimulus and does not include the question, which in my opinion is a flaw in the model.

Besides, not including features that relate to the scientific language, such as technical vocabulary and the use of everyday language, makes the model unsuitable for assessing demands of science tests, let alone do that across languages. Variable 6 ‘Semantic match between task and text’ is also relevant to the language variable. But a simple investigation of the extent of overlap between the task and the text does not tell us much about the demands brought about by the semantics employed in each language. Hence, just like the other two models of comparing judgement of demands, this model is also not suitable for comparing judgements of demands of language versions of the same assessment. The next section recapitulates all the issues discussed earlier about the different methods explored in judging demands.

8.4.4 Challenges to judging demands

In the previous sections of this chapter, I reviewed the various methods and models employed in rating demands to shed light on the practical and methodological problems associated with them. These are summarised below.

Collecting judgements is a time-consuming and costly endeavour especially when non-examiners, such students and teachers are involved, as they require more time to become familiar with the assessment materials and the procedures (Pollitt et al., 2007). In most of the studies reviewed earlier, examiners were involved in judging the demands of the

questions. Some studies also investigated the ability of examiners to judge the difficulty of items, that is, to predict the success rate of students on particular questions (Wolf et al., 1995; Meyer, 2003). Wolf et al. (1995) asked a panel of examiners to rate the difficulty of 30 maths questions administered to over 300 students. Significant yet moderate correlations (0.4) were found between examiners' estimation of success rates and the empirical difficulty of the question. This study suggested that examiners are not capable of judging the order of difficulty of items for students.

Similar findings were reached by Meyer (2003) who investigated the ability of examiners to judge the success rate of A level economics multiple choice questions for students who would be awarded a failing grade (grade E) on the examination. Meyer also found moderate correlations and a difference between examiners' estimation of success rates and the actual empirical difficulty of the items.

Arguments in favour of using non-examiners contend that students and teachers could have a better understanding of what makes an examination demanding (Pollitt et al., 2007). Students, being novices at the subjects being assessed, do not necessarily perceive demands in the same way expert judges do. Wood and Pollitt (2006) carried out construct elicitation interviews with A level students and confirmed that students can provide very useful insight into the way demands are judged. Student-generated data was coherent and their perceptions of what makes a question demanding differed significantly from those of experts. On the contrary, Graesser et al. (1999) asked university students to rate the demands of several surveys. They found that university students were not capable of giving thorough explanations of what was problematic in some survey questions.

Hopfenbeck and Maul (2011) carried out cognitive interviews (also referred to as 'think-aloud' protocols) with twenty-two Norwegian students after they had sat PISA 2006. In

cognitive interviews, interviewees respond to interviewers' questions while answering the survey items. This approach allows the examination of the cognitive processes taking place while students are answering survey. The authors used this method to investigate the validity of the Norwegian version of PISA attitudinal scales. While this method allows in-depth analysis of the demands questions place on students, Hopfenbeck and Maul (2011) acknowledged several of its limitations. The method is costly; besides, actual cognitive processes underlying students' responses might not be well reflected. Students tend to behave differently under assessment conditions than in interview contexts. In low-stake assessments, students may select answers randomly without reading the questions. In cognitive interviews, they are compelled to read them carefully before providing an answer.

Likewise, it has been argued that teachers who deal with students' challenges more closely on a regular basis may be better at judging demands (Pollitt et al., 2007); however, no published paper has sought teachers rating of demands to this date. Nevertheless, results of a study using a panel of twenty-three teachers to estimate the difficulty of items in a science test suggested that there was a low accuracy in teachers' judgements (Impara & Plake, 1998).

Rating overall demand has been reported to be problematic (Edwards & Adams, 2002) and not informative and most studies consist of rating components of demands. Adding up components without sound justifications is problematic especially when studies often report interactions between the different components. Some studies using the CRAS framework have computed arithmetic averages of the four CRAS scales. This method could be acceptable if the averages are only taken as indications of exam papers with more or less problematic features. No further interpretations of results should be made.

There is a concern that breaking down 'overall demand' into smaller components generates an interaction amongst these components thereby inflating the rating of overall demand (e.g. McLone & Patrick, 1990; Jones, 1993). If there is no interaction then there is likely an averaging effect towards the middle category (Fowles, 1995) which brushes off significant differences between exams. Moreover, problems with understanding the construct statements have been reported especially when it comes to some scales of CRAS. Promoting a shared understanding through discussion amongst judges has been recommended (Greator et al., 2002) although this procedure does not necessarily eliminate the problem (Fearnley, 2000).

The reliability and the validity of data collected has been criticised by Adams and Pinot de Moira (2000). Even when judges shared the same meaning for each construct statement, it did not mean they applied the same values to the categories within the scale. Moreover, Pollitt et al. (2007) warned that the use of the construct elicitation technique often led to a large variance between examiners' ratings and hence inter-rater reliability was likely to be low. Besides, intra-judge consistency has not been formally investigated so interpretation of results of these studies should be made with caution.

It is worth highlighting that methods of rating demands are still qualitative and attempts to quantify them failed to predict difficulty. Studies managed to detect significant differences in the nature of demands but they have not been yet successful at measuring their size or educational significance (Pollitt et al., 2007). In addition, with the exception of the last model, the Lumley et al. Task Difficulty model, all the other frameworks have been mainly employed to compare high-stakes examinations in the UK; that is, in the same country. The components do not highlight factors relevant to cross-cultural comparisons. Language is summarised in one or two components of demands and is not extensively assessed. In

the following section, I look closely at few aspects of PISA science items and show that none of the models previously described would be able to detect the differential demands implicated by the language of assessment in language versions of PISA science 2006 test.

While robust methodology seems to be the main complication in gauging the difficulty and demands associated with an item, a more conceptual challenge underlie the failure in doing so. It is related to the confusion between the concept of difficulty and that of demand. There seems to be little consensus on what each of these terms entails. Pollitt et al.'s (2007) distinction is a first step to separate the two concepts. However, their attempt does not clear the ambiguity entirely.

Pollitt et al.'s definition of item difficulty is confined to the classical test theory paradigm where item difficulty is the proportion of people getting an item correct, in other words, the p -value. However, under the IRT paradigm, item difficulty has a slightly different meaning. Item difficulty is calibrated in relation to person ability estimates in 1-PL models, to item discrimination estimates in 2-PL models and to pseudo-guessing estimates in 3-PL models.

Moreover, Pollitt et al.'s conception of demands as qualitative only does not fully cater for educators' and examiners' needs which may legitimately include the mapping of demands on a scale. Indeed the conception of demands as qualitative contradicts the CRAS model Pollitt et al. (1998) proposed which consists of four scales of demands. It is thus not surprising that methods relying on judgment of demands yield very low reliability given the ambiguity surrounding the concept of demands. Future research should sharpen the distinction between the concepts of difficulty and demands of a task or item while bearing in mind the various paradigms underlying assessment and the utility of the concepts to the users.

8.5 Problems with survey questions

Research carried out on survey methodology is of immense relevance to judging demands placed by questions in PISA assessments. Indeed, PISA assessments are described by OECD as cognitive item surveys (OECD, 1999). Apart from the educational content of the subject-specific PISA assessments, PISA surveys are very similar to common surveys such as census surveys in that participants have very vague expectations about the type and content of the questions and the assessment is of low-stakes to the individuals.

It is true that surveys and educational assessments differ on various aspects, namely in terms of content, participants' expectations and the level of stakes they represent to the participant. However, they both require participants to process questions, understand them and provide answers. Hence, the cognitive processes and the reading comprehension mechanisms involved in reading questions are similar in both educational assessments and survey questions to a great extent.

Questions imposing high cognitive load have been associated with significant levels of response error (Bless et al., 1992; Knäuper et al., 1997; Velez & Ashworth, 2007). Questions which are hard to understand lead to variability in respondents' answers and hence poor validity and reliability of the tests (Fowler & Cannell, 1996; Groves 1989; Lessler & Kalsbeek, 1993; Schober & Conrad 1997). Moreover, literature reviewed by Lenzner et al. (2010) found that when faced with questions that are difficult to comprehend, respondents tend to either give wrong answers or less optimal and merely satisfactory ones. They are also less likely to complete the entire survey.

Despite the negative impact of low comprehensibility on the quality of data in surveys, available guidelines for writing good survey questions are not of great assistance. Lenzner

et al. (2010) criticised these guidelines for being too vague and for suggesting avoiding the obvious, such as long or complex questions and unfamiliar terms. They do not indicate what makes a question complex and when it is considered too long. They do not specify what terms are considered less familiar. The guidelines rely heavily on subjective interpretation of survey developers.

Recent research on survey questions explored cognitive science theories and advances in discourse processing and reading comprehension. These areas of research are not subject-specific and can help understand language aspects involved in cognitive processing of questions irrespective of their content. In this section, I highlight main problems with writing survey questions identified in the literature and relevant to judging demands in educational assessments.

Graesser and colleagues (1996) carried out studies in preparation for the development of a computer programme called the Question Understanding Aid (QUAID). QUAID aims to diagnose and detect problems in questions. This aspect will be further examined in the Discussion chapter. In the course of this research, trained researchers with different expertise in psycholinguistics, cognitive science and discourse processes analysed questions from a wide variety of survey corpora. These included information surveys, applications to graduate school, dentist intake forms, etc. The analyses led to the identification of twelve problematic features in questions (see Table 8.3, items 1 to 12). Graesser et al. (1996) found that the first five variables – (1) unfamiliar technical term, (2) vague or imprecise predicate or relative term, (3) vague or ambiguous noun phrase, (4) complex syntax; (5) working memory overload – occurred the most in survey questions.

Lenzner et al. (2010) proposed two additional variables to Graesser et al.'s (1996) first five variables. These are presented in Table 8.3 as item (13) low syntactic redundancy and (14)

bridging inferences. Besides, Lenzner et al. (2010) referred to variable (1) as word frequency rather than term familiarity. Their rationale was based on findings that familiarity was a vague and a non-objective measure of word difficulty (Colombo et al., 2006; Rayner et al, 2012) and that word frequency and familiarity are highly correlated anyway (Balota et al., 2006). Based on a web experiment comparing two versions of similar questions, Lenzner et al. (2010) measured cognitive effort using response times; that is, the time respondents need to process questions before they generate an answer. Results suggested that cognitive burden was mainly caused by complex syntax and working memory overload. Their analyses led them also to recommend avoiding the use of acronyms, vague quantification terms, left-embedded syntactic structures, quantitative mental calculations and hypothetical questions.

Table 8.3: Problems with survey questions (based on Graesser et al., 1996 and Lenzner et al., 2010)

	Problem	Description
1.	Unfamiliar technical term or low frequency term	There is a word or expression that very few respondents would know the meaning of.
2.	Vague or imprecise predicate or relative term	The values of a predicate (i.e. main verb, adjective, or adverb) are not specified on an underlying continuum.
3.	Vague or ambiguous noun phrase	The referent of a noun phrase, noun, or pronoun is unclear or ambiguous
4.	Complex syntax	The grammatical composition is embedded, dense, structurally ambiguous, or not well formed syntactically.
5.	Working memory overload	Words, phrases, or clauses impose a high load on immediate memory.
6.	Misleading or incorrect presupposition	The truth value of a presupposed proposition is false or inapplicable.
7.	Unclear question category	It is difficult to determine what class of question is being asked.
8.	Amalgamation of more than one question category	The question may be assigned too two or more different classes of questions.
9.	Unclear question purpose	The respondent may not know why the question is being asked.
10.	Mismatch between question category and answer option	The question invites one set of answer options that is different from the question options of the questionnaire.
11.	Difficult to access specific or generic knowledge	A typical respondent would have difficulty recalling the information requested in the questions.
12.	Respondent unlikely to know answer (no information source)	A typical respondent would not know the information requested in the question.
13.	Low syntactic redundancy	The grammatical structure of a sentence is not predictable.
14.	Bridging inferences	The question forces respondents to draw inferences in order to bridge information and establish coherence.

Table 8.3 is by no means an exhaustive list of problematic features in survey questions. These factors have been shown to impose cognitive burden or impede comprehension. A very important aspect of this list is that it comprises variables that are grounded in theories

of cognitive science, discourse processing and reading comprehension. Cognitive science offers a number of relevant theories that allow a better judgment of demands in educational assessments. In the next section, I explore cognitive load theory and highlight the extent to which this theory can be helpful in examining question demands.

8.6 Judging demands using cognitive load theory

In the previous sections, I reviewed the main approaches and models reported in the literature for judging demands placed by items on examinees in educational settings. It is perhaps worth noting that all the reviewed methods have been published as conference papers or reports released by exam boards or the exam regulator in the UK (Qualification and Curriculum Agency (QCA) at that time). None of these were published as articles in peer-reviewed journals, which could be an indication of the immature state of the field.

Nevertheless, more rigorous and conclusive studies on judging cognitive demands have been carried out in cognitive psychology. Indeed, this area has been explored by Graesser et al., (1996) to identify variables making survey questions hard to comprehend. Various models and theories have been proposed to understand the cognitive demands involved while an individual is learning or completing a given task. It is beyond the scope of this thesis to review all these models in depth. However, given the challenges of judging demands imposed on candidates by question papers, it is worth turning to cognitive psychology and incorporating some relevant theories and techniques in educational assessment. Of particular significance are the cognitive load theory (CLT: Paas & van Merriënboer, 1994a; Sweller et al., 1998) and the effect of incongruence on working memory (WM). In this section, I first provide a brief overview of the CLT and then

highlight the role of working memory. Later, I discuss the implications of cognitive load on educational assessment.

8.6.1 Cognitive Load Theory

CLT enables the design of instructional methods that optimise learning and performance because of the theory's central concern with humans' restricted cognitive processing capacity (Paas et al., 2003). CLT relies on a cognitive structure composed of a limited WM which, as suggested by Baddeley (1986), includes two processing units for visual/spatial and auditory/verbal information. The first is called the visuo-spatial sketchpad and the second one is referred to as the phonological loop. According to CLT, efficient and successful learning happens when individuals' minds organise and interpret information into cognitive frameworks called schemas. Schemas are saved in the long term memory and can be largely automated with increased exposure to the same information (Paas et al., 2003). This chunking of information into cognitive schemas and their subsequent automation relieves much of the burden on the WM, thereby circumventing its constraints.

A strong correlation exists between performance on measures of WM capacity and performance on higher-order cognitive tasks such as reading comprehension, learning and reasoning (Daneman & Carpenter, 1980). Effective learning and performance are highly affected by the amount of WM resources involved in a particular task. Learning and performance are optimal when enough cognitive resources are allocated. Conversely, learning fails and performance deteriorates in cases of cognitive underload or cognitive overload; that is, when task demands fall behind or exceed the available cognitive capacity. Hence, in educational assessments, it is crucial to maintain cognitive load balanced in a way to keep examinees positively challenged without frustrating them. A question worth asking then is how to strike the right balance of cognitive load? Paas & van

Merriënboer (1994a) proposed a model which described the complexity of the cognitive load variable and provides a conceptual framework of how it can be measured and hence manipulated. The model is described below.

According to Paas & van Merriënboer (1994a), cognitive load comprises two dimensions – a causal dimension and an assessment dimension. The causal dimension conveys the interaction between the task characteristics and the participant's attributes. Task characteristics include, amongst several others, task format, task complexity and time pressure. Participants' attributes include mainly their age and level of expertise. These variables are particularly important as they define what Pass et al. (2003) call the 'intrinsic cognitive load.' That is, the cognitive load defined by the nature of the task and the level of expertise of the learner or examinee. Intrinsic cognitive load cannot be manipulated by instructional designers.

The assessment dimension consists of three measures – mental load, mental effort and performance. Mental load refers to a form of cognitive load that arises from the interaction between task and participants' characteristics. It can be considered as an *a priori* estimate of cognitive load. Understanding the nature of the interaction and being able to measure mental load could be instrumental in estimating cognitive demands prior to test administration. Mental effort is the allocated cognitive capacity to cope with the demands of a given task. It is considered as an estimate of the actual cognitive load. Performance refers to the participant's achievement on a given task and is measured in terms of number of correct or incorrect responses or time spent on task. Performance in this model is what Pollitt and colleagues (2007) refer to as item or task difficulty although according to them, 'time spent on a task would be considered as a measure of demand, not difficulty.'

The three measures – mental load, mental effort and performance – interact to keep performance optimal (Paas & van Merriënboer, 1993). When faced with complex tasks with a high mental load, people often invest additional mental effort to maintain high performance. However, if the mental load was too high, mental effort can reach a ceiling effect leading to performance degradation. The key feature to a successful design of instructional tasks is to keep mental load manageable. Test developers cannot manipulate intrinsic cognitive load because it is inherent in the interaction between a given task and a given examinee.

According to CLT, two other forms of cognitive load exist– extraneous cognitive load and germane cognitive load (Paas et al., 2003). Extraneous cognitive load, also referred to as ineffective load, is the load exceeding intrinsic cognitive load and is usually generated by poorly designed instruction. Germane cognitive load, also referred to as effective load, relates to processes that enable the building and automation of mental schemas. Based on CLT, total cognitive load is the sum of the three forms of cognitive load. For successful learning or performance to occur, overall cognitive load should remain within working memory limits and germane cognitive load should be enhanced while extraneous cognitive load should be minimised.

8.6.2 Effect of incongruence on the working memory (WM)

Incongruence between two cognitive processes can overwhelm the WM and impair selective attention. A classic experiment to illustrate this effect is the Stroop effect (Stroop, 1935). While there are now many versions of this experiment, the original series of experiments consisted of presenting participants with a series of words of colours (red, blue, orange, green, black, etc.) printed in different ink colours. In the first presentation set, there is congruence between the ink colour and the meaning of the word (for instance, the

word 'blue' is displayed in blue ink). In the second set, there is incongruence between the ink colour of the words and the meaning of the word (that is, the word "blue" appears in a different colour, for example yellow). Participants are asked to name the ink colour of words (that is, they should not read the words). Researchers record the reaction time (that is, the time it takes participants to give the correct response) and the error rate.

Results showed that reaction time and error rate were significantly higher in the incongruent presentation set. Early explanations of the Stroop effect created by the interference of the target task (naming colours) with the distractor task (reading words) were related to the 'automaticity hypothesis' (Stroop, 1935; LaBerge & Samuels, 1974; Shiffrin & Schneider, 1977). According to this view, colour naming is more taxing and requires more attention than word recognition which is considered to be rather automatic. Word reading thus impeded the more strenuous task of colour naming (Luo, 1999).

More recent explanations refute the automaticity hypothesis and are based on the perceptual load theory, which contends that the processing of a stimulus (in this case, naming a colour) can be reduced by increasing the perceptual load of a task in the working memory (in this case providing conflicting visual distractor, a non-matching word) (Lavie, 1995; Rees, Frith & Lavie, 1997; Lavie et al., 2004). When the WM capacity is overloaded, in this case with words which are incongruent with the colour in which they are presented, the WM becomes less capable of inhibiting distractors and attentional selection is impaired. Experiments carried out by Kim and colleagues (2005) suggest that negative interference only occurs when tasks share the same limited resources of WM. This is in line with multiple resource theory (Wickens, 2002; 2008) and the fact that WM is not only about memory but also about attention (Engle, 2002).

The case of incongruence is of prime relevance in educational assessment. Often the way tasks or questions are presented creates incongruence situations. Incongruence cannot be spotted by any of the models discussed in earlier sections and can be unpredictable across languages. In the next section, several examples are provided from PISA science items showing incongruence situations generated because of test adaptation. However, before examining these examples, below is a brief summary of the main methods used in cognitive psychology to judge cognitive load.

8.6.3 Techniques used to measure cognitive load

Various measurement techniques have been developed to measure cognitive load. Most of the studies employed empirical approaches. The most common of these is participants' indicating the amount of cognitive load experienced when completing a task on various rating scales. The scales include variables such as mental effort, fatigue, frustration which are highly correlated (Nygren, 1991). These methods are popular because of their practicality, sensitivity and relatively low cost compared to other methods despite the lack of evidence around their internal consistency (Paas et al., 2003).

Some methods relied on physiological responses such as heart-rate variability which proved to be unreliable (e.g. Paas & van Merriënboer; 1994b) or pupillary dilation which proved to yield more satisfactory results (Beatty and Lucero-Wagoner, 2000). It would be interesting to see whether some physiology-based techniques such as pupillary dilation could provide reliable information for judging item demands in educational assessment. Other methods involved the measurement of performance (reaction time, accuracy, error rate) on a secondary task carried out simultaneously with a primary task. Analytical approaches using expert judgement and mathematical models have been developed yet seldom used. While some of the techniques used in cognitive psychology could be

appealing for their potential higher reliability (e.g. the physiology-based techniques); however, these methods can only measure cognitive load in terms of ‘amount’ but cannot point to the problematic feature in the questions or task that could be causing a given demand. International assessment is in need of more thorough judgement methods which can handle the impact of language and culture variables on demands. Research in cognitive psychology might not provide the most appropriate techniques to judge demands; however, any models for describing cognitive demands in educational assessment should incorporate insights from this area of research.

8.7 A closer look at some of the released items

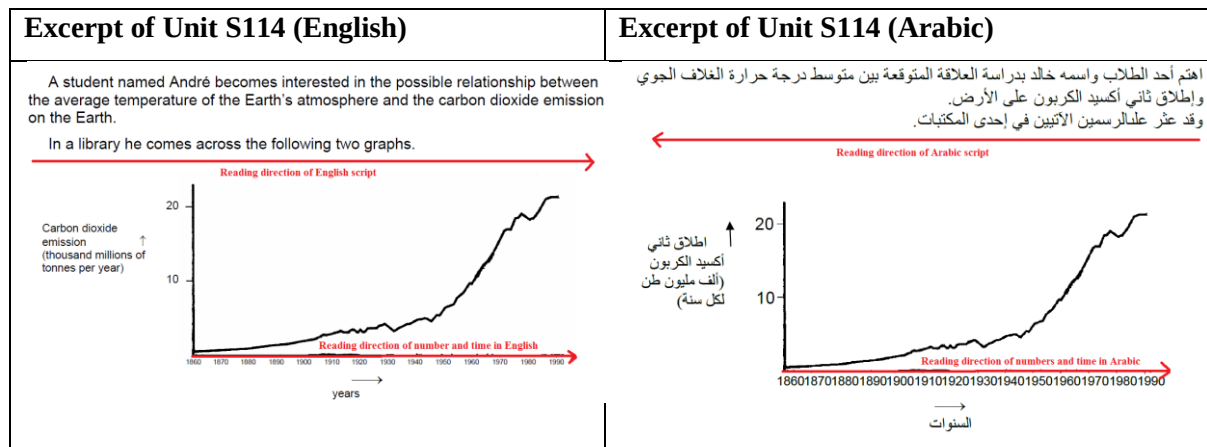
In the previous section, I provided an overview of cognitive load theory and how this area of research can provide new ways of understanding and judging demands in educational assessment. In addition to the various concerns raised in these reports about these methods and summarised in this chapter, I argued earlier that none of these methods can handle comparability studies of international tests. In this section, I illustrate this point with examples from PISA science items and highlight some of the variables that are inherent in the language and that would not be picked up by the methods reviewed earlier. I draw on research carried out in the field of psychology on cognitive burden and the Stroop effect.

8.7.1 Semiotics incongruence

Unit S114 (See Appendix to 3.1) of PISA science 2006 includes questions requiring students to read graphs. Graphs are semiotic tools with which scientists communicate data and mastering them is a part of acquiring science literacy (Chapter 2). I have discussed in Chapter 3 that the skill of reading graphs is not necessarily emphasised to the same extent

by age 15 in all curricula. Besides, reading graphs is not equally demanding in all languages. For instance, in English and French, the time line moves from left to right and hence matches the direction of reading in both languages. In Arabic, this is not the case. Time lines and numbers are read from left to right while the language script is read from right to left. See Figure 8.3 below.

Figure 8.3: English and Arabic versions of graphs in PISA unit S114



Source: www.oecd.org/pisa/38709385.pdf.

This aspect of the Arabic language has nothing to do with the quality of translation and is inherent in the language. The additional cognitive demand that is placed upon examinees because of the incongruence between aspects of the items cannot be picked up by generic models that do not focus on the language variables in an international setting. This incongruence, not originally present in the source version, is likely to result in an additional WM load for Arabic examinees as similar WM resources are needed in both tasks, reading a graph and reading the text. Another form of incongruence resulting from test adaptation is described below.

8.7.2 Abbreviation incongruence

The translation of chemical elements and compounds can be less intuitive from one language to another because the symbols are universal and use Latin letters. For instance,

the English symbol for lead is 'Pb'. There is incongruence between the word 'lead' and the symbol of the chemical element 'Pb' which is an abbreviation of the Latin origin of the word 'Plumbum' meaning soft metals. Translating an item containing this symbol into French eliminates the incongruence because the French term for lead is '*plomb*'. This makes the item less demanding in French.

Other examples of elements or compounds could have the opposite effect. For instance the chemical symbol for the nitrogen element is 'N' in English which creates no incongruence between the symbol and the name of the element. However, the French commonly used term for nitrogen is 'azote' (from Greek meaning 'no life'). Hence, translating an item from English to French with the element nitrogen will make the element more demanding in French because of an incongruence that was not originally there in the English version of the item.

A similar case can be described in PISA science 2006 items where one of the units, Unit S447 Sunscreens, mentions the chemical compound zinc oxide. The chemical formula of the compound is ZnO. The chemical formula matches exactly the English term 'zinc oxide' and is the reverse of the French term '*oxyde de zinc*'. The French term might only create very little additional demands if any. However, the Jordanian Arabic version uses the term '*أكسيد الخارصين*' [*oxeed al kharaseen*] which on the one hand, is a different script to the universal chemical formula ZnO and on the other hand, does not include the sound /z/ but the guttural sound /x/. It is interesting that the Jordanian translators chose this term instead of a more intuitive alternative '*أكسيد الزنك*' [*oxeed al zinc*] which uses the term zinc, also used in Arabic.

One explanation could be that [*oxeed al kharaseen*] is the term found in the Jordanian curriculum and the one that students are most familiar with. However, it does not eliminate

the incongruence between the name of the chemical compound and the Latin formula ZnO. Similar problems are likely to occur in Chinese and Japanese languages. Again, this incongruence is likely to increase WM load as both concurrent tasks involve reading; either reading the name of the compound or its chemical formula.

8.7.3 Interference between everyday language and scientific language

I have discussed earlier (Section 8.4.1.b) that students are challenged by terms which are commonly used in everyday language yet have a specific meaning in science. This interference could be understood as interference between the different schemas constructed. Earlier in this chapter (Section 8.4.1.b), I gave the example of the meaning of the term ‘force’ in science and its common use in informal settings. When examinees are faced with such words in science items, one can hypothesise that the WM could get overloaded as it attempts to activate the right schemas relevant to science and inhibit the other more mundane meaning acting as a distractor.

Although most languages may have the overlap between everyday language and scientific language yet this overlap may not be on exactly the same terms. In other words, translating a term like ‘weight’ which has a double meaning in English into another target language may not be problematic as the target language may use a science-specific term for weight. The opposite is also true. Some scientific terms which have everyday usage may be translated into other languages which use that same term in more informal settings.

Below is an example from PISA science 2006 that illustrates the point. Item S447Q02 is a multiple choice question that requires participants to identify the role of substances in a set of experiments. D is the key of this question ‘*Mineral oil and zinc oxide are both reference substances.*’ Reference substances, as their name implies, are used as a point of

comparison. The translation of ‘reference substance’ into French leads to a very similar expression ‘*substance de reference*’ which again gives a clue about the role of the substance. However, the translation into Arabic becomes a bit problematic as the technical term is ‘مادة ضابطة’ [madda dabita] the common everyday meaning of the term [dabit] is ‘officer or army general’.

The connotation of this word as a reference substance is less common in the everyday language and the role of the substance becomes less obvious than in English or in French. It is an empirical question whether this term will increase cognitive load because it activates different sets of schemas. However, this example illustrates again that adaptation can lead to differential cognitive demands.

Unit S447 (Sunscreens), item S447Q02

Which one of these statements is a scientific description of the role of the mineral oil and the zinc oxide in comparing the effectiveness of the sunscreens?

- A Mineral oil and zinc oxide are both factors being tested.
- B Mineral oil is a factor being tested and zinc oxide is a reference substance.
- C Mineral oil is a reference substance and zinc oxide is a factor being tested.
- D Mineral oil and zinc oxide are both reference substances

Source: www.oecd.org/pisa/38709385.pdf.

8.7.4 Unfamiliar technical acronyms

In the Sunscreen unit (Unit S447), the stimulus mentions the term Sun Protection Factor and its acronym SPF. The first problem with this item is that it includes an acronym and as previously mentioned, it is recommended to avoid the use of acronyms in questions (Lenzner et al., 2010). Another problem relates to how successfully acronyms can be

translated. The French translation of SPF is the equivalent term *facteur de protection solaire (FPS)*. The translation into Arabic gives ‘معامل حماية’ [moua’amel himaya] and similarly the acronym makes use of the first letter of each word (ح.م.) [mim ha’].

I carried out an online search to check whether the Arabic equivalent of ‘Sun Protection Factor’ or its acronym are as well known in the Arabic culture as they are for English-speaking people and noticed that the term is not frequently used and the acronym does not exist. One should see this in the culture’s context. Arabic-speaking countries are located in warm sunny areas and awareness about protection from the sun is less salient. That is why acronyms like SPF are not of big importance. Besides, searching advertisement of sunscreens sold in the Arab world resulted in finding that the term SPF is used with the Latin letters even when advertisements are written in Arabic. Another advertisement included the term ‘معامل حماية’ along with SPF to clarify what the Arabic term means as SPF is more familiar to Arabic-speakers. The question worth asking here is whether by using the Arabic equivalent for SPF, the demands are increased for Arabic-speaking examinees compared to their English-speaking or French-speaking counterparts who are probably more familiar with the term and its acronym.

Figure 8.4: Advertisements of cosmetic product in Arabic with the acronyms UVA and SPF in Latin letter²³.



The previous examples are only illustrations of aspects that were not captured by the approaches and models for judging demands reviewed in sections 8.3 and 8.4. Different cognitive demands could be triggered by different languages and these could be sometimes difficult to avoid because they are rooted in the nature of the language. The universality of the chemical formula and the incongruence it ensues in languages not using a Latin alphabet is not within the remit of translation or adaptation techniques.

CONCLUSION

Until the early nineties, exploring demands placed on examinees when taking a test relied on *ad hoc* methods such as questionnaires asking for their views on relative demands of exam papers (Houston, 1981; Massey, 1979). The studies lacked a robust methodology and concerns about judges' understanding of questionnaire statements were raised (McLone & Patrick; 1990). More systematic approaches followed the introduction of the National Curriculum (age 5 to 16) in Britain (Jones, 1993; Ratcliffe, 1994; Stobart et al., 1994). These consisted of more defined procedures for training judges and rating demands.

²³ <http://bashrety.tv/ar/index.php>.

Similar issues arose with these methods in addition to potential bias in the rating (Jones, 1997). The application of Kelly's (1955) construct elicitation technique in some studies (e.g. Adams & Pinot de Moira, 2000; Fearnley, 2000; Gray, 2000; Greateorex et al., 2002; 2003; Pritchard et al., 2000) allowed the rating of a greater number of exam features. However, most of these were not describing demands (e.g. Fearnley, 2000; Greateorex et al., 2002) and results were contradicted by other studies using the same data but different statistical analyses (e.g. Baird, 1999). Similar problems were highlighted in the methods adopted in studies examining the source of DIF in language versions of PISA (Arffman, 2012; Huang et al., 2014). In these studies, an additional challenge is associated with the bias introduced from the use of monolingual and multilingual judgment designs.

The Question Answering Process (QAP) model proposed by Pollitt and Ahmed (1999) described the stages involved in when an examinee is responding to a question. This model suggested that the most critical stage was the first one when examinees were reading and understanding a question. A number of features making questions easier or more difficult have been identified. These were referred to as sources of easiness (SOEs) and sources of difficulty (SODs). While these features may have highlighted aspects making questions more or less understandable, studies adopting this approach failed at manipulating demands so as to predict difficulty. One main issue with this approach is confounding the qualitative nature of demands with the more statistical aspect of item difficulty (Fisher-Hoch et al., 1997).

Another model was proposed by the same main author and his colleagues, the CRAS model (Pollitt et al., 1998). This model focused on cognitive demands and allowed rating on four scales: complexity, resources, abstractness and strategy. This model was more

successful at describing demands but like other approaches, it suffered from inconsistency of judgements and the inability to predict item difficulty.

A more recent framework describing task difficulty was proposed to predict the difficulty of PISA reading tests (Lumley et al., 2012). This framework seemed more promising as it yielded more reliable results and seemed to work well with non-English versions of PISA. However, no other studies have been published using this model and it was not clear whether it can be used with science tests.

Given the lack of suitable models and approaches to judge demands of questions in exam papers across languages, I turned to cognitive load theory which examined the way cognitive load is influenced by factors that burden the WM. Some of this burden is caused by incongruence between the meaning of words and the way they are represented leading to attention impairment. In this chapter, I provided examples from released PISA science items illustrating how the adaptation of tests can increase cognitive load because of the nature of the target language. No amount of quality control could change the fact that reading graphs in Arabic and reading Arabic texts occur in opposite direction causing incongruence.

Research in cognitive theory seems to be a promising way forward to investigate demands entailed in language versions of the same educational assessment. Efforts in bridging this area with educational assessment can help test developers improve the quality of assessments. However, language being intrinsic to the construct measured forms a confounding variable that is difficult to isolate and control for. This makes equivalence of constructs across languages very challenging. Comparing language versions of the same assessment becomes akin to comparing apples and oranges. We can go as far as saying they are both types of fruit yet with no genetic engineering involved, apples and oranges

will always taste, smell and feel differently. Language versions of the same test can be at best as close as the red colour is to the orange colour. In the Discussion Chapter (Chapter 9), I argue that the way forward for judging demands in educational assessment, especially in the international context, lies within an interdisciplinary endeavour and the joint forces of experts in educational assessments, cognitive psychology and computational linguistics.

Chapter 9

Discussion

The fact that language is a bio-cultural hybrid is its most important property, and a key to understanding our own place in nature.

Evans and Levinson (2008)

INTRODUCTION

This thesis investigated whether different languages, because of their idiosyncrasies, imposed different linguistic and cognitive demands that could eventually be reflected as different proportions of students answering a given item correctly. It attempted to answer the overarching research question: are language versions of the same science assessment comparable in terms of question difficulty and demands. After over four years of research and grappling with many fundamental questions of measurement, psychometrics, language philosophy and other, I reached the frustrating conclusion that this comparison is conceptually and methodologically unachievable.

Why unachievable? In a nutshell, this is because language is a confounding variable in what we are assessing. It is not a variable that can be neutralised especially when comparing language versions of the same assessment administered to people from different cultures and holding different world views. Moreover, the methodologies available until this date are inadequate. Techniques adopted for developing and translating tests rely heavily on subjective expert judgement and methods used in analysing the results rely on assumptions that do not necessarily stand in educational settings. Besides, DIF techniques do not allow the detection of DIF when the test is entirely biased in one direction. All items could be harder in a given language version of a test yet this difference would not be manifested as DIF. Ensuring equivalence at test level becomes less certain and cross-

cultural comparisons at item and at test level become less possible. In this chapter, I develop my claims further and support them with conceptual arguments and empirical evidence from the literature.

The previous two paragraphs capture briefly my DPhil journey. Despite all the challenges, this journey was far from being depressing; it was very enriching and exciting. My interest in this research area stemmed from my personal experience as a science teacher in the Lebanese system where science education and assessment occur in English and French, not in Arabic, (see Chapter 1), the country's national language. Soon later, I realised that this question was also raised at a larger international scale with the growing importance of international assessments such as PISA, TIMSS and PIRLS. PISA was used in this research as a context for the conceptual arguments I advanced. I chose PISA because it is currently the most influential international assessment and has instigated major educational reforms in some countries such as Germany and Japan (Ertl, 2006; Schleicher, 2009; Takayama, 2008). However, the claims are to a large extent valid to any cross-cultural assessment attempting to use language versions of the same test to compare performance of individuals across countries.

My confidence in generalising more broadly the validity of this argument is rooted in the essence of the research questions I raised and the methodology I adopted. In this thesis, I used an in-depth critical evaluation of the design, methodology and results of PISA. The critical evaluation was informed by literature in various fields: educational assessment, psychometrics, philosophy of measurement, cognitive science, philosophy of language, and computational linguistics. I also carried out analyses using the Rasch model to illustrate empirically many of the claims I advanced. Additionally, I used released PISA items to demonstrate many of the controversies around translation and adaptation of items.

My critique of the design, techniques and results of PISA should not be considered as an *ad hominem* criticism of PISA itself. It is rather an evaluation of the available methodology and some of the claims based on PISA which did not stand on robust conceptual grounds. Indeed, I see great value in comparing educational systems across countries for the sake of exploring country differences and reflecting on one's educational reality. However, cross-cultural comparisons gave rise to heated debates and controversies, the most famous of these in recent times being the Kreiner and Christensen (2014) critique of the use of the Rasch model in PISA and the strong evidence of differential item functioning in their study. My research has tapped into several of these debates making my journey more complex yet more thrilling. In this final chapter, I summarise and discuss my main claims under three headings reflecting three major controversies jeopardising the comparability of language versions of the same assessment.

Controversy 1: Language is embedded in what we are testing.

Controversy 2: The statistical methods and models used in international assessments do not capture the complexity of educational constructs.

Controversy 3: Available methods are not suitable for detecting invalidity of international comparisons.

I follow these three sections with a last section where I propose a way forward and the exploration of computational linguistic techniques. Before getting into the juicy details of the chapter, I owe the reader a clarification of the quotations that I have translated literally into English at the beginning of Chapters 1 to 6. The first section of the chapter provides an explanation in English of each of the sayings opening these chapters.

9.1 Translating meaning of popular idioms

- Chapter 1 Arabic saying: [ya mestaajel wa'ef ta ellak] 'يا مستعجل وقف تلاك'

Literally meaning: ‘Hey you in a hurry! Stop so that I tell you.’ This expression is used when people are delayed by various incidents (such as not finding the car keys and then being caught in a traffic jam on the way to an important appointment).

- Chapter 2 Greek saying: *Τα μάτια σου δεκατέσσερα*. [Ta mátia sou dekatéssera]

In English, it translates literally to ‘Your eyes fourteen.’ This expression is used to ask someone to be careful, to keep their eyes open at all times.

- Chapter 3 French saying: *‘Le diable bat sa femme et marie sa fille.’*

In English this literally means ‘The devil beats his wife and marries his daughter.’ This is an expression used during a rain shower while it is sunny and a rainbow appears.

- Chapter 4 Dutch saying: *‘Ga toch fietsen!’*

In English it literally means ‘Ride a bike!’ This exclamation is used when in frustration with someone, like ‘I don’t believe you, stop bothering me!’ I cannot but highlight the cultural aspect of the bike and the strong association it has with the Dutch culture; the Netherlands being widely known for its cycling-friendly roads.

- Chapter 5 Norwegian saying: *‘Å vere midt i smørauget’* (Nynorsk)

This translates literally into ‘to be in the middle of the butter melting in the porridge.’ It is an expression meaning to be in a very favourable place or situation.

- Chapter 6 German saying: *‘Es ist zum Mäuse melken.’*

This translates literally into: ‘It is to milk mice!’ It is said when a situation is very complicated and frustrating.

These quotes were purposefully placed at the beginning of each chapter to keep the reader reflecting on the translation challenges. The sayings reflect how delicate language is and

how simple it might be to translate words but how challenging it is to translate meaning. Guidelines for translation of tests do warn against literal translation; however, the literature is full of empirical evidence that shows that despite following strict guidelines and carrying out thorough revisions of translated and adapted versions of texts and assessments, the impact of translation and adaption cannot be entirely neutralised. Indeed, *No amount of painstaking translation will quite eradicate the Englishness of thought which underlies it* (Jones, 1988, p.179). The following sections tackle the issue by highlighting three controversies from technical, methodological and conceptual perspectives.

9.2 Controversy 1: Language is embedded in what we are testing

Constructs in educational assessment are built based on curricula or conceptual frameworks of a certain discipline. PISA assessments are also developed accordingly; their science tests are based on a scientific literacy framework (OECD, 2009b). While at first glance the framework did not appear to assess students' language proficiency, delving into the literature on scientific literacy and examining closely the PISA science competencies and proficiency levels as well as the science items painted a totally different picture. The literature highlights the importance of reading, writing and communication in acquiring scientific literacy; an aspect that Norris and Phillips (2003) referred to as the fundamental sense of scientific literacy as opposed to its derived sense which focuses on the scientific knowledge. Research in science education also describes the peculiarities of the language of science which are characterised by its syntactical complexity and high density in technical terms.

A careful inspection of PISA science competencies revealed that PISA science tests measured the fundamental as well as the derived sense of scientific literacy. This was also clearly reflected in the description of the proficiency levels of PISA science which

associated high proficiency in scientific literacy with advanced reading and writing skills in science as well as a solid command of the scientific language. Also, looking at the reading and writing demands placed on learners by the length of the items and the expected responses confirmed the importance of language in PISA science assessments. All this evidence suggested that language was part of the construct in science assessments and more so in PISA science tests even if it was not intended by the PISA consortium.

If language is part of the science construct in PISA, this has tremendous implications on the comparability of constructs across countries most of which use different languages. The challenge of having equivalent constructs in international assessments has been already signalled in the literature (Hambleton et al., 2005) and has been demonstrated empirically (Ercikan & Koh, 2005). Ercikan and Koh compared the English and French versions of the same TIMSS 1995 test across various countries (Canada, England, France and USA) and investigated DIF using a 2-PL IRT approach. This research examined the comparability of PISA constructs across three languages with data from Jordan and the Arabic version of PISA 2006 being for the first time explored. Grisay et al. (2007) recommended that more studies should focus on equivalence of PISA versions of Middle East countries as in these appear to have suffered most from lesser level of comparability.

Moreover, I applied a new approach to investigate differential item functioning at item and at test level. Oliveri and Ercikan (2011) investigated DIF at item and at test levels in PISA problem-solving 2003 tests using latent class analysis. I used the Hagquist and Andrich (2004) approach to explore DIF at item level based on a two-way ANOVA of response residuals and designed a new method to investigate DIF at test level. Neither of these approaches was previously adopted to explore DIF in international assessments.

My empirical work has contributed new ways of exploring DIF in international assessment. However, my main contribution in the field lies in the critical evaluation I carried out and the conceptual framing of the research question. I used the linguistic relativism principle to argue that because language shapes our thinking and is embedded in assessment be it intentionally a part of the construct or be it not, constructs can never be equivalent across language and hence across cultures. Despite all rigorous translation and adaptation techniques, I used various examples to show that languages due to their idiosyncrasies place different linguistic and cognitive demands. These differences can often lead to differential item difficulties and this was supported by empirical evidence based on the comparisons of PISA, TIMSS and PIRLS data across countries (Ercikan & Koh, 2005; Grisay & Monseur, 2007; Grisay et al., 2007; Hauger & Sireci, 2008; Huang et al., 2014; Kreiner & Christensen, 2014; Le, 2009; Oliveri & Ercikan, 2011; Sandilands et al., 2013; Wu & Ercikan, 2006; Xie & Wilson, 2008).

In Chapter 3, I reviewed the literature on adaptation and translation of instruments. I listed the various methods and techniques employed in creating language versions of the same instrument (translation, adaptation, application options). Later, I outlined the various sources of bias encountered in the process of developing language versions of the same assessment targeting different cultures.

PISA adapts its assessments from two source versions (English and French) using a double translation technique in an effort to decentre the assessment, that is, to make it less culturally biased (OECD, 2009b). The quality control process includes various rounds of revision carried out by trained translators who are also often experts in the subject matter assessed. Despite the thorough translation process, empirical evidence reveals the emergence of bias across language versions of the same test (Grisay & Monseur, 2007;

Grisay et al., 2007; Huang et al., 2014; Kreiner & Christensen, 2014; Le, 2009; Oliveri & Ercikan, 2011; Xie & Wilson, 2008).

International standards and guidelines have been developed to ensure high quality and equivalence of adapted tests (ITC, 2001). Nevertheless, the final verdict about the quality of equivalence in international assessments is made subjectively by panels of experts (OECD, 2009b). Empirical evidence points to the low agreement amongst judges (Graesser et al., 2006; Pollitt et al., 2007). According to Hambleton (2005), irrespective of whether backward, forward or double translation is involved in the adaptation process, verifying the quality of the adapted test ideally involves either:

- an expert (or a team of experts) in the subject matter who is also competent in the source and target languages and has a background in assessment. He/she compares the two versions and evaluates the quality of the translation in relation to specific criteria.
- a panel of raters assessing the quality of the adapted version using scores on some criteria. Then a form of agreement is reached on amendments that need to be made to ensure equivalence of the two versions.

As discussed in Chapter 5, PISA makes use of expert judgement to judge the equivalence of the national version with one or both source versions. Empirical evidence suggests that these methods are unreliable and lack strong theoretical grounding (Graesser et al., 2006; Pollitt et al., 2007). Indeed, despite the rigorous quality control exerted on the translation and adaptation processes of international tests, some items have been reported to function differently in different countries (e.g. Bonnet, 2002; Ercikan & Koh, 2005; Grisay & Monseur, 2007; Grisay et al., 2007; Hauger & Sireci, 2008; Huang et al., 2014; Kreiner & Christensen, 2014; Le, 2009; Oliveri & Ercikan, 2011; Sandilands et al., 2013; Wu & Ercikan, 2006; Xie & Wilson, 2008).

In this section, I recap the argument that language is part of what we are measuring in science assessments and that rigorous translation and adaptation methods fail to neutralise the language variable. I argue that this is due to the intimate relationship between language and thinking. In the next section, I discuss how statistical methods used in international assessment fail to cope with the complexity of the educational assessment reality and the inextricable nature of the language element within educational constructs, namely in the science discipline.

9.3 Controversy 2: The statistical methods and models used in international assessments do not capture the complexity of educational constructs

Central data analyses in PISA are carried out using the Rasch model (OECD, 2009a, b). This model is attractive thanks to its simplicity and its capacity to map individuals' abilities on a latent trait on a single scale of item difficulty. The model is distinguished by its property of invariance (Rasch, 1961). In educational assessment, this implies that if the data fit the model, the comparison between two individuals is irrespective of the item used for the comparison and similarly, the comparison between two items is independent of the person used for the comparison (Rasch, 1961). At first glance, the Rasch model looks ideal for establishing comparisons across countries. However, Kreiner and Christensen (2014) have criticised the use of the model and the violation of its requirements. This debate brings up various contentions discussed in the following sections about the purpose of psychometric models, the nature of the construct, the importance of data fit and the assumptions on which these models lie.

9.3.1 Measuring or modelling?

In the field of psychometrics, the choice of model is often motivated by deep epistemological stances. According to proponents of Rasch modelling, the model subscribes to the requirements of fundamental measurement in the physical sciences (Wright, 1997; Bond & Fox, 2001). In Chapter 4, I distanced myself from such strong statements. I made clear that I did not claim to be carrying out scientific measurement in social sciences. Indeed, there is a huge contention in the field of psychology, led by Joel Michell, Denny Borsboom and David Andrich, about whether statistical methods used, psychometrics particularly, aim at measuring a given ability or modelling it; i.e. describing it. The conflict was compared to the Kuhnian view of incompatible paradigms (Andrich, 2004).

The main controversy revolves around the nature of the object being measured or described; in this case, human intellectual abilities. Are these quantifiable? Can we prove that a given intellectual ability, such as intelligence, is a quantity as defined in science? Until now, such a proof has not been provided for any psychological (and educational) construct and it looks like the opposing parties are talking past each other (Andrich, 2004). Borsboom (2005), in his book *Measuring the Mind*, relates this ‘sterile’ dialogue to the different ontologies and epistemologies. On one end of the continuum, realists are interested in measuring constructs; on the opposite end relativists are only interested in describing constructs. This relates to how individuals perceive constructs: is a psychological construct a real entity or is it the product of the researcher’s imagination? Different epistemologies lead to different choices of research models.

9.3.2 Data fit to the model

The use of Rasch modelling was fiercely criticised in PISA (Goldstein et al., 2007; Kreiner & Christensen, 2014). Kreiner and Christensen (2014) attacked the use of the Rasch model for various reasons; one of which being the great misfit of data to the model. In analyses I carried out on PISA 2006 Science data of UK, France and Jordan (Chapter 6), significant misfit was detected in 11 out of 103 items. Rasch modellers tend to find that 15% misfit is within the limits of what they consider acceptable (Adams & Wright, 1994). Kreiner and Christensen's (2014) analysis of a subset of reading items of PISA 2006 showed that all but three items misfit the model and all items showed DIF. Goldstein (2004) criticised against the prioritisation of data fit over the educational purpose of items. According to many, educational objectives inherently violate many of the model's assumptions and particularly that of unidimensionality. The next sections focus on the issues associated with the assumptions underlying psychometric models.

9.3.3 The assumptions underlying IRT models

The assumptions of the IRT models have been the source of much reservation for the use of psychometric techniques in educational assessment. Baird and Black (2013) criticised the use of psychometrics in public examinations in Britain because of their implications on the design and reliability of the assessments. They maintained that the requirements of psychometrics often conflicted with the educational, cultural and political concerns of a given nation. Goldstein (1979) argued that often educational constructs are aligned with specific curricular objectives. Items targeting these constructs may violate requirements of psychometric models to serve more eminent educational goals. According to Baird and Black (2013), psychometrics seems to be solving someone else's problems.

9.3.3.a Local independence

For instance, local independence is one of these assumptions. In the analysis of PISA data I carried out, five pairs of items showed response dependence with two pairs showing a relatively high magnitude of dependence (0.7 logits). These items clearly violated the requirement of local independence on which the Rasch model stands. According to proponents of Rasch modelling, the anomaly is in the questions and it may well be. However, Baird and Black (2013) maintained that it is not possible to ask higher order thinking questions without questions with some level of response dependence.

9.3.3.b Unidimensionality

Another main aspect that seems to escape psychometricians' attention is the complexity of educational reality. PISA uses the Rasch model which measures the ability of students on a unidimensional scale defined by items of varying difficulty. IRT models used in international assessments, including the Rasch model, rely on the assumption that the test measures a single construct. This statement has been empirically challenged. Indeed, Goldstein et al. (2007) used various multilevel modelling and structural equation techniques in the analysis of PISA reading 2000 data in England and France. Their study suggested that PISA Reading tests have at least two dimensions. Multidimensional IRT models do exist yet they are not widely used in international assessments.

In the principle component analysis I carried out on variance residuals, the first two components accounted for less than 6% of the total residual variance. This suggested that the requirement of unidimensionality was not violated. However, these findings conflicted with the results of the DIF analyses which I will discuss shortly given that DIF could be a manifestation of multidimensionality (Marais & Andrich, 2008b). Besides, a quick look at the science items in PISA suggests that the items do measure different aspects of the

scientific literacy construct. For instance, the science item on greenhouse effect (see Chapter 8, Section 8.7.1) requires participants to answer a question based on a graph. Students are asked to establish a causal relationship and formulate a clear answer after having successfully read two graphs and the text in the accompanying stimulus. Hence, in this item, at least five different elements are measured: the scientific knowledge (the greenhouse effect), reading a scientific text, reading a graph, writing a scientific answer and establishing causal relationships. These reflect well the complexity of the scientific literacy construct that targets the fundamental and the derived dimensions of scientific literacy (Norris & Phillips, 2003).

In Chapter 4, I argued based on Stenner et al. (2008; 2009) and Andrich (2014) that scientific literacy is a unidimensional thick index variable consisting of three thinner variables (or indices): knowledge *about* science, knowledge *of* science and scientific language. Unidimensional psychometric models used in international assessments fail to capture the complexity of the scientific literacy construct. They are incapable of providing information about each index measured within an index variable. This might not serve the educational purpose of monitoring achievement on every aspect of the construct. A mastery of the greenhouse effect concept does not necessarily mean being capable of reading a graph or being able to employ scientific language properly. In addition, these models fail to indicate the source of bias in measurement. This point is further elaborated in Controversy 3 below.

9.4 Controversy 3: Available methods are not suitable for detecting invalidity of international comparisons.

Common methods used in comparing language versions of the same assessment include differential item functioning techniques. In Chapter 7, I reported results of a DIF analysis

carried out on PISA Science 2006 data of three countries (UK, France and Jordan). I used Hagquist and Andrich's (2004) approach of detecting DIF. The analysis suggested that 23 out of 103 items showed DIF; that is 22% of the items. This finding was consistent with previous studies carried out on international large scale assessment (e.g. Ercikan & Koh, 2005; Grisay & Monseur, 2007; Huang et al., 2014; Oliveri & Ercikan, 2011).

I also investigated whether the DIF items advantaged or disadvantaged a particular country. Results of repeated-measures ANOVA suggested that despite the presence of DIF at item level, the bias balanced out at test level making differences in means of DIF items not significant. Often, DIF is a manifestation of multidimensionality (Marais & Andrich, 2008b). However, the balancing of the DIF across countries was consistent with the PCA results in Chapter 6 which suggested that the test was unidimensional at test level at a certain level of precision. Balancing DIF across groups is a useful technique to get around inevitable bias in cross-cultural assessments. The question worth asking however is whether the balancing out of DIF is a robust reason to consider comparisons valid despite the presence of bias at item level.

A major problem with DIF is that it has several limitations. First, DIF analyses only detect inconsistencies in the behaviour of a given item in comparison with all other items. DIF is thus not an absolute measure and cannot detect bias if all items were advantaging or disadvantaging a single country. Besides, DIF does not allow the identification of sources leading to its emergence. This is particularly problematic in international assessments where countless variables distinguish the groups compared: cultural differences, language, curriculum, nature of educational systems, school types, to name only a few. DIF could be caused by one of these factors or a combination of them. Hence, the field lacks the proper

statistical techniques to identify whether language is causing DIF. This makes the answer to my research question methodologically very challenging.

DIF is reflected in significant differences of item difficulty across groups. In Chapter 8, I adopted Pollitt et al.'s (2007) distinction between difficulty and demand; the former referring to the empirical proportion of individuals getting an item correct while the second relating to the cognitive process placed on individuals during assessment.

A great number of empirical studies have been carried out on comparability of test demands. Most of these studies have been carried out on British examinations to maintain standards across examination boards, years or subjects. Early approaches lacked rigour and were based on *ad hoc* methods (Houston, 1981; Massey, 1979). Later studies which followed the introduction of the National Curriculum in Britain became more systematic and relied on trained judges rating demands (Jones, 1993; Ratcliffe, 1994; Stobart et al., 1994). Kelly's (1955) construct elicitation technique became popular in the 2000s (e.g. Adams & Pinot de Moira, 2000; Fearnley, 2000; Gray, 2000; Greaux et al., 2002; 2003; Pritchard et al., 2000). However, all these studies suffered various methodological flaws, especially an inability to identify what places demands on students when answering questions (e.g. Fearnley, 2000; Greaux et al., 2002) and low reliability of results (e.g. Baird, 1999).

Attempts to compare item demands and the impact of translation in international assessments were carried out on PISA (Arffman, 2012; Huang et al., 2014). Despite the systematic processes adopted, the methods suffered design flaws, namely the introduction of bias in the judgement when adopting monolingual and multilingual reviewers (Hambleton, 2005).

Several models were proposed for judging demands and comparing them across different exam papers. In Chapter 8, I reviewed three models: QAP (Pollitt & Ahmed, 1999), CRAS model (Pollitt et al., 1998) and Lumley et al.'s (2012) task difficulty framework. The models are definitely a theoretical step forward towards a better understanding and judgement of test demands. However, none of them has been used in a published peer reviewed journal article. This could be an indication of their poor reliability and robustness. Besides, none of them allowed an extensive exploration of the language variable and how it affects demands. Except for Lumley's task difficulty framework, the other two models have been developed to serve purposes of comparability of assessments in one country (the UK).

In Chapter 8, I reviewed literature in cognitive psychology to suggest aspects in questions that impose demands that were not identified in educational assessment research such as incongruence caused by how questions are presented. These aspects could be manifested to different degrees in different languages. I illustrated my arguments with several examples taken from PISA science 2006 to show how the previously reviewed models could not detect some type of demands and hence would not be helpful when comparing demands across language versions of the same assessment. Luckily, there seems to be a way forward, which requires however joint forces of researchers in several fields: disciplinary experts, educational assessment, cognitive psychology and computational linguistics. The next section discusses this proposition and gives an example of an available online facility for detecting demanding features in survey questions. This facility could be adapted to serve educational assessment purposes.

9.5 The way forward: Exploring cognitive science and computational linguistics

In the 1980s, successful machine translation was almost unimaginable within a generation. A widely used example in natural language processing was the machine translation of the biblical statement “*The spirit is willing but, the flesh is weak*” (St Matthew, 26:41) into Russian. Its back translation from Russian gave ‘*The vodka is good, but the meat is rotten*’ (Rich, 1984, p.341). However, typing the English version today into a widely used machine translation programme such as Google Translate gives decent translations in French and Arabic. Figure 9.1 compares the Google Translation and the actual biblical translations. A bilingual (or trilingual) person can notice the similarity. Actually, the term ‘flesh’ would not translate readily into ‘الجسد’ [body in Arabic] without a deeper level of translation embedded in the programme.

Figure 9.1: Quality of machine translation in 2015

	Google Translate	Biblical version
French	<i>L’esprit est bien disposé mais la chair est faible.</i>	<i>‘L’esprit est ardent, mais la chair est faible.’</i>
Arabic	روح على استعداد ، ولكن الجسد ضعيف.	‘أما الروح فنشيط، وأما الجسد فضعيف.’

A lot of progress has been witnessed in the areas of natural language processing and computational linguistics. This section describes a tool called QUAID developed at the Institute of Intelligent Systems (IIS) at the University of Memphis. This tool could be adapted to be used in educational assessment settings and developed further to cope with more languages. QUAID has the potential to constitute the basis for more objective methods to compare demands imposed by language in educational assessments.

The Question Understanding Aid (QUAID) is an online facility provided by IIS to systematically identify problems with questions based on theoretical grounds (Graesser et al., 2000). The development of QUAID was informed by research in cognitive psychology

and survey methodology which analysed the process of answering questions. QUAID is only concerned with the interpretation stage of the process as it is one of the major sources of measurement error which threaten the reliability and validity of answers to questions (Fowler & Cannell, 1996; Groves 1989; Lessler & Kalsbeek 1993; Schober & Conrad 1997). QUAID diagnoses problems associated with the interpretation of questions.

In Chapter 8, I outlined 14 main problems suggested to challenge understanding of questions (Graesser et al., 2006; Lenzner et al., 2010) with the first five being the most critical (Graesser et al., 1996; Graesser et al., 1999). QUAID currently diagnoses problems related to the first five variables in English: (1) words that are unfamiliar to many adults; (2) unclear relative terms (i.e. verbs, adjectives, adverbs); (3) vague or ambiguous noun phrases; (4) questions with complex syntax, and (5) questions that overload the working memory. The programme critiques each question for potential comprehension difficulty at various levels of language, discourse and world knowledge.

QUAID has great potential for assisting question setters in detecting features that may pose problems in questions. Some of these problems may be missed by survey methodologists because of fatigue or lack of proper training. QUAID's output was compared with ratings carried out by language experts on a large corpus of over 500 questions of different surveys. The aforementioned classes of problems were used as a basis for the ratings. Despite false alarms detected by QUAID (that is, erroneously signalling problematic features), QUAID was capable of detecting problems relevant to the five classes.

No published studies have used QUAID in educational settings. Further research could look into adapting QUAID to flag problematic features in educational questions. To compare demands posed by these problematic features, QUAID could be extended to incorporate corpora of assessments in different languages. QUAID could be also updated

with more recent research carried out in the field by Graesser and McNemara (2011) who built another computer programme (Coh-metrix) to evaluate discourse complexity using a multilevel framework of discourse comprehension. The discussion of the multilevel framework proposed by Graesser and McNemara is beyond the scope of this thesis but will definitely be of great use in investigating a way forward for looking at language demands in assessment across cultures.

9.6 Limitations

This research was faced with several limitations:

- This research analysed data from only one PISA cycle (2006) and only one subject (science). Mathematics and readings are likely to highlight additional and possibly more or less serious challenges in international comparisons. Nevertheless, the critical evaluation approach adopted in this thesis allows the broadening of many of the arguments to other PISA cycles and to other subjects.
- The initial sample of the three countries was reduced to 4,000 in each country (12,000 in total). This is in line with procedures adopted when dealing with large sample size. Tests of fit in Rasch are significance tests which are highly sensitive to sample size. This means that part of the information (even if minimal) has been lost because of this procedure.
- As discussed earlier, DIF does not identify the sources of its emergence and hence, it is difficult to explain it, especially with no access to the items.
- Despite tremendous efforts in acquiring PISA non-released items, access to these items was never granted. This impacted the extent to which DIF items could be analysed qualitatively. Only three out of the 23 DIF items were publically available.

9.7 Ethical issues

This research consisted of a secondary analysis of PISA data publically available online. In this sense, ethical issues are not a concern. However, deeper ethical issues are associated with the nature of the research question. Equivalence of tests and comparability of scores are directly linked to matters of fairness and equity. These questions are of prime importance within the broader education and scientific community.

9.8 Concluding remarks

This thesis attempted to answer the following research question: are language versions of the same science assessment comparable in terms of language difficulty and demands? A fundamental problem arises when attempting to compare cognitive abilities across nations. Rigorous guidelines for translating and adapting tests are available; yet this does not solve the underlying deeper conceptual issue of language and culture diversity. As soon as tests are translated, they become different at a certain level. An independent yardstick used as a reference is needed; however, once again, this reference would suffer from the same problem. There is no way to know the extent to which translation impacts item difficulty and demand. Quantitative and qualitative methodologies are incapable of handling the inextricable nature of language in psychological and educational constructs.

This thesis compared language versions of the same PISA test of three countries UK, France and Jordan. While several studies investigated the comparability of English and French versions of international assessments (Ercikan & Koh; 2005; Oliveri & Ercikan, 2011), no previous research explored the Arabic version of PISA science tests.

In this thesis, I argued that language is part of the science construct irrespective of the intention of the question setter. Because language shapes thinking, different languages

impose differential demands because of their idiosyncrasies. These are often manifested in differential item difficulties which could be detected as DIF. I investigated DIF at item and at test level. DIF at item level was explored by using the Hagquist and Andrich (2004) method which has not been applied in international assessment contexts yet. DIF at test level was investigated while examining significant differences in mean item locations of DIF items across countries and comparing mean person estimates before and after splitting items in each country. Both approaches have not been used in any of the previous studies investigating DIF at item or at test level in international assessment.

DIF techniques allow flagging bias in items using fairly simple procedures. Nevertheless, they do not allow detecting the sources of bias. Unfortunately, until today, the field lacks suitable methodologies to deal with these challenges despite all efforts of thoroughness involved in the process. One practical way of dealing with DIF is to ensure that it is balanced across groups. Educational assessment is in essence a practical discipline where pragmatics often motivates decisions. Given that DIF is not an absolute measure and is a function of the test and groups involved, balancing can only occur post-hoc if not a fortunate empirical result.

The critique I made in this thesis was not in any way an opportunity to discredit genuine efforts put in building international assessments. It should be seen as a reflection on what has been done so far and the ways with which we can address the challenges international comparisons pose. After all, science proceeds through critique and debate. This thesis does not only add to the field of educational assessment but also contributes to broader issues of translation of psychological and educational tests. In this thesis, I bridged different fields otherwise disparate. Although Art Graesser's work linked computational linguistics with cognitive science and theories of learning, no previous work bridged educational

assessment, test adaptation and translation, cognitive load theory and computational linguistics. This stems from a strong confidence that a way forward can only be build based on joint interdisciplinary efforts across these fields.

Reference List

- Abedi, J. (2004) The No Child Left Behind Act and English Language Learners: Assessment and Accountability Issues. *Educational Researcher*, 33 (4), pp. 4-14.
- Adams, R.J., Wilson, M. and Wang, W. (1997) The Multidimensional Random Coefficients Multinomial Logit Model. *Applied Psychological Measurement*, 21 (1), pp. 1-23.
- Adams, R.J. and Wright, B.D. (1994) When does misfit make a difference? In *Objective measurement: Theory into practice. Volume 2.*, ed. M. Wilson, Norwood, New Jersey: Ablex., pp. 244-270.
- Adams, R.M. and Pinot de Moira, A. (2000) *A comparability study in GCSE French including parts of the Scottish Standard grade examination. A study based on the summer 1999 examination. Review of question paper demand, cross-moderation study and statistical analysis of results.* Organised by Welsh Joint Education Committee and Assessment and Qualifications Alliance on behalf of the Joint Forum for the GCSE and GCE.
- Ahmed, A. and Pollitt, A. (1999) Curriculum demands and question difficulty. *Paper presented at IAEA Conference*, Slovenia
- Allalouf, A. and Ben-Shakhar, G. (1998) The effect of coaching on the predictive validity of scholastic aptitude tests. *Journal of Educational Measurement*, 35 (1), pp. 31-47.
- American Educational Research Association, American Psychological Association and National Council on Measurement in Education (1999) *Standard for Educational and Psychological Testing*. Washington, DC.: American Educational Research Association.
- American Educational Research Association, American Psychological Association and National Council on Measurement in Education (1985) *Standard for Educational and Psychological Testing*. Washington, DC: American Psychological Association.
- and *Evaluation: An International Journal on Theory and Practice*, 16 (6), pp. 451-470.
- Andersen, E.B. (1973) A goodness of fit test for the Rasch model. *Psychometrika*, 38, pp. 123-140.
- Andersen, E.B. (1977) Sufficient statistics and latent trait models. *Psychometrika*, 42, pp. 69-81.
- Andersen, E.B. (1995) Residual analysis in the polytomous Rasch model. *Psychometrika*, 60, pp. 375-393.
- Andersen, E.B. (2002) Residual diagrams based on a remarkably simple result concerning the variances of maximum likelihood estimators. *Journal of Educational and Behavioral Statistics*, 27 (1), pp. 19-30.
- Anderson, C.W. (1999) Inscriptions and science learning. *Journal of Research in Science Teaching*, 36 (9), pp. 973-974.
- Andrich, D. (1978) A rating formulation for ordered response categories. *Psychometrika*, 43 (4), pp. 561-574.
- Andrich, D. (1982) An index of person separation in latent trait theory, the traditional KR.20 index and the Guttman scale response pattern. *Education Research and Perspectives*, 9, pp.95-104.
- Andrich, D. (1985) A latent trait model for items with response dependencies: Implications for test construction and analysis. In *Test Design.*, ed. S.E. Embretson, New York, NY: Academic Press, pp. 245-275.
- Andrich, D. (1988) *Rasch models for measurement*. Newbury Park, CA: Sage.
- Andrich, D. (1989) Distinction between assumptions and requirements in measurement in the social sciences. *Proceedings of the XXIVth International Congress of Psychology*, 4, 7-16. *Mathematical and Theoretical Systems*, eds. J.A. Keats, R. Taft, R.A. Heath & S.H. Lovibond, B.V. North Holland: Elsevier Science Publications.
- Andrich, D. (1995) Distinctive and incompatible properties of two common classes of IRT models for graded responses. *Applied Psychological Measurement*, 19 (1), pp. 101-119.
- Andrich, D. (2006) *On the fractal dimension of social measurement*. Unpublished paper ed. Murdoch University, Perth, Australia: Social Measurement Laboratory.
- Andrich, D. (2014) A structure of index and causal variables. *Rasch Measurement Transactions*, 28 (3), pp. 1475-1477.
- Andrich, D. and Hagquist, C. (2012) Real and artificial differential item functioning. *Journal of Educational and Behavioral Statistics*, 37 (3), pp. 387-416.
- Andrich, D. and Hagquist, C. (2015) Real and Artificial Differential Item Functioning in Polytomous Items. *Educational and Psychological Measurement*, 75(2), pp.185-207.

- Andrich, D. and Kreiner, S. (2010) Quantifying response dependence between two dichotomous items using the Rasch model. *Applied Psychological Measurement*, 34 (3), pp. 181-192.
- Andrich, D., Marais, I. and Humphreys, S. (2012) Using a theorem by Andersen and the dichotomous Rasch model to assess the presence of random guessing in multiple choice items. *Journal of Educational and Behavioral Statistics*, 37 (3), pp. 417-442.
- Andrich, D., Sheridan, B.E. and Luo, G. (2009) *RUMM2030: Rasch Unidimensional Models for Measurement*. Perth, Western Australia: RUMM Laboratory.
- Andrich, D.A. (2004) Controversy and the Rasch model: A characteristic of incompatible paradigms? . *Medical Care*, 42 (1), pp. 1-15.
- Angoff, W.H. (1993) Perspectives on differential item functioning methodology In *Differential Item Functioning.*, eds. W. Holland and H. Wainer, Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers, pp. 3-23.
- Arffman, I. (2012) International education studies: Increasing their linguistic comparability by developing judgmental reviews. *International Scholarly Research Network*, [Online]. 2012 , pp. 30/08/14. Available from: <<http://www.hindawi.com/journals/isrn/>> (Accessed 30/08/14).
- Astolfi, J. (1997) *L'Erreur, un Outil pour Enseigner*. Paris: ESF.
- Ausubel, D.P., Novak, J.D. and Hanesien, H. (1978) *Educational Psychology: A cognitive View*. 2nd ed. New York: Holt, Rinehart and Winston.
- Baddeley, A.D. (1986) *Working Memory*. Oxford: Oxford University Press.
- Baird, J. (1999) *Regression analysis of review outcomes*. Paper presented at a seminar held by the Research Committee of the Joint Council for General Qualifications, Manchester.
- Baird, J. and Black, P. (2013) Test theories, educational priorities and reliability of public examinations in England. *Research Papers in Education*, 28 (1), pp. 5-21.
- Baird, J., Ahmed, A., Hopfenbeck, T., Brown, C. and Elliott, V. (2013) *Research Evidence Relating to Proposals for Reform of the GCSE*. OUCEA Report.
- Baird, J., Hopfenbeck, T.N., Newton, P., Stobart, G. and Steen-Utheim, A.T. (2014) *Assessment and Learning: State of the Field Review*. Oslo: Norwegian Knowledge Centre for Education.
- Balota, D.A., Yap, M.J. and Cortese, M.J. (2006) Visual word recognition: The journey from features to meaning (a travel update). In *Handbook of Psycholinguistics.*, eds. M.J. Traxler and M.A. Gernsbacher, Vol.2 ed. Amsterdam: Elsevier, pp. 285-375.
- Beatty, J. and Lucer-Wagoner, B. (2000) The pupillary system. In *Handbook of Psychophysiology.*, eds. J.T. Cacioppo L.G. Tassinary and G.G. Berntson, 2nd ed. Cambridge MA: Cambridge University Press, pp. 142-162.
- Bland, J.M. and Altman, D.G. (1995) Multiple significance tests: the Bonferroni method. *British Medical Journal*, 310, pp. 170.
- Bless, H., Böhner, G., Hild, T. and Schwarz, N. (1992) Asking difficult questions: Task complexity increases the impact of response alternatives . *European Journal of Social Psychology*, 22, pp. 309-312.
- Bloom, B.S. (1956) *Taxonomy of educational objectives: The classification of educational goals by a committee of college and university examiners. Handbook I: Cognitive domain*. New York: Longmans, Green.
- Bond, T.G. and Fox, C.M. (2001) *Applying the Rasch model: fundamental measurement in social sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Bonnet, G. (2002) Reflections in a Critical Eye: On the pitfalls of international assessment. *Assessment in Education: Principles, Policy and Practice*, 9 (3), pp. 387-399.
- Boroditsky, L. (2001) Does language shape thought?: Mandarin and English speakers' conceptions of time. *Cognitive Psychology*, 43, pp. 1-22.
- Boroditsky, L. and Gaby, A. (2010) Remembrances of times East: Absolute spatial representations of time in an Australian aboriginal community. *Psychological Science*, 21, pp. 1635-1639.
- Borsboom, D. (2005) *Measuring the Mind: Conceptual Issues in Contemporary Psychometrics*. Cambridge: Cambridge University Press.
- 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. . OECD Publishing.
- Briggs, D.C., Alonzo, A.C., Schwab, C. and Wilson, M. (2006) Diagnostic Assessment With Ordered Multiple-Choice Items. *Educational Assessment*, 11 (1), pp. 33-63.
- Brodgen, H.E. (1977) The rasch model, the law of comparative judgment and additive conjoint measurement. *Psychometrika*, 42 (4), pp. 631-634.

- Bunting, B.P. and Mooney, E. (2001) The effects of practice and coaching on test results for educational selection at eleven years of age. *Educational Psychology: An International Journal of Experimental Educational Psychology*, 21 (3), pp. 243-253.
- Bybee, R. (1997) *Achieving Scientific Literacy: From Purposes to Practices*. Portsmouth, NH: Heinemann.
- Bybee, R. and McCrae, B. (2011) Scientific literacy and student attitudes: Perspectives from PISA 2006 science. *International Journal of Science Education*, 33 (1), pp. 7-26.
- Bybee, R., McCrae, B. and Laurie, R. (2009) PISA 2006: An assessment of scientific literacy. *Journal of Research in Science Teaching*, 46 (8), pp. 865-883.
- Carroll, J.B. (ed.) (1956) *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf*. Cambridge, MA: MIT Press.
- Chavkin, L. (1997) Readability and reading ease revisited: State-adopted science textbooks. *The Clearing House*, 70, pp. 151-154.
- Chomsky, N. (1986) *Knowledge of Language: its Nature, Origin and Use*. New York.: Praeger.
- Cohen, K. (1962) The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, 65, pp. 145-153.
- Colombo, L., Pasini, M. and Balota, D.A. (2006) Dissociating the influence of familiarity and meaningfulness from word frequency in naming and lexical decision performance. *Memory & Cognition*, 34, pp. 1312-1324.
- Cronbach, L.J. and Meehl, P.E. (1955) Construct validity in psychological tests. *Psychological Bulletin*, 52 (4), pp. 281-302.
- Daneman, M. and Carpenter, P.A. (1980) Individual differences in WM and reading. *Journal of Verbal Learning and Verbal Behavior*, 19, pp. 450-466.
- D'Arcy, J. (1997) *Comparability studies between modular and non-modular syllabuses in GCE Advanced level biology, English literature and mathematics in the 1996 summer examinations*. Standing Committee on Research on Behalf of the Joint Forum for the GCSE and GCE.
- DeBoer, G.E. (2000) Scientific literacy: Another look at its historical and contemporary meanings and its relationship to science education reform. *Journal of Research in Science Teaching*, 37 (6), pp. 582-601.
- Dhillon, D. (2003) *Predictive Models of Question Difficulty: A Critical Review of the Literature*. Manchester: AQA Centre for Education Research and Policy.
- Dobzhansky, T. (1973) Nothing in biology makes sense except in light of evolution. *American Biology Teacher*, 35, pp. 127-129.
- Domínguez, M., Vieira, M. and Vidal, J. (2012) The impact of the Programme for International Student Assessment on academic journals. *Assessment in Education: Principles, Policy & Practice*, 19 (4), pp. 393-409.
- Dorans, N.J., Schmitt, A.P. and Bleistein, C.A. (1992) The Standardization Approach to Assessing Comprehensive Differential Item Functioning. *Journal of Educational Measurement*, 29 (4), pp. 309-319.
- Duncan, O.D. (1984) Rasch measurement: Further examples and discussion. In *Surveying Subjective Phenomena*, eds. C.F. Turner and E. Martin, New York: Russel Sage Foundation, pp. 367-401.
- Dunn, O.J. (1961) Multiple comparisons among means. *Journal of the American Statistical Association*, 56, pp. 52-64.
- Edwards, E. and Adams, R. (2002) A comparability study in GCE AS geography including parts of the Scottish Higher grade examination. A study based on the summer 2001 examination. Organised by the Welsh Joint Education Committee on behalf of the Joint Council for General Qualifications.
- Edwards, E. and Adams, R. (2003) *A comparability study in GCE Advanced level geography including the Scottish Advanced Higher grade examination. A study based on the summer 2002 examination*. Organised by the Welsh Joint Education Committee on behalf of the Joint Council for General Qualifications.
- Edwards, J. and Dall'Alba, G. (1981) Development of a scale of cognitive demand for analysis of printed secondary science materials. *Research in Science Education*, 11, pp. 158-170.
- Eivers, E. (2010) PISA: Issues in implementation and interpretation. *The Irish Journal of Education / Iris Eireannach an Oideachais*, 38, pp. 94-118.
- Eklöf, H. (2010) Skill and will: test-taking motivation and assessment quality. *Assessment in Education: Principles, Policy and Practice*, 17 (4), pp. 345-356.
- Eklöf, H., Hopfenbeck, T.N. and Kjaernsli, M. (2012) *Hva vet vi om elevers testmotivasjon? Erfaringer fra internasjonale og nasjonale undersøkelser i Norge og Sverige (What do we know about students' test motivation? Experiences of international and national tests in Norway and*

- Sweden). In *Kvalitet i Norsk Skole. Internasjonale og Nasjonale Undersøkelser av Læringsutbytte og Undervisning. (Quality in the Norwegian School. International and National Tests of Learning Outcomes and Teaching)*., eds. T.N. Hopfenbeck M. Kjaernsli and R.V. Olsen, Oslo, Norway: Oslo, Universitetsforlaget.
- El Masri, Y. and Vlaardingerbroek, B. (2010) Science textbook readability in Lebanon: A comparison between anglophone and francophone learning milieux. *Mediterranean Journal of Educational Studies*, 15 (1), pp. 109-124.
- Engelhard, G.J. (1992) Historical views of invariance: Evidence from the measurement theories of Thorndike, Thurstone, and Rasch. *Educational and Psychological Measurement*, 52, pp. 275-291.
- Engle, R.W. (2002) Working memory capacity as executive attention. *Current Directions in Psychological Science*, 11 (1), pp. 19-23.
- Ercikan, K. (2002) Disentangling sources of differential item functioning in multilanguage assessments. *International Journal of Testing*, 2 (3-4), pp. 199-215.
- Ercikan, K. and Koh, K. (2005) Examining the construct comparability of the English and French versions of TIMSS. *International Journal of Testing*, 5 (1), pp. 23-35.
- Ercikan, K., Gierl, M.J., McCreith, T., Phan, G. and Koh, K. (2004) Comparability of bilingual versions of assessments: Sources of incomparability of English and French versions of Canada's national achievement tests. *Applied Measurement in Education*, 17(3), pp.301-321.
- Ertl, H. (2006) Educational standards and the changing discourse on education: The reception and consequences of the PISA study in Germany . *Oxford Review of Education*, 32, pp. 619-634.
- Evans, B.F. and Pierce, G.E. (1982) *Report of a GCE inter-board study in German at the Advanced level 1980*. Cardiff: Welsh Joint Education Committee.
- Evans, N. and Levinson, S.C. (2009) The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32 (5), pp. 329-492.
- Fang, Z. (2006) The language demands of science reading in middle school. *International Journal of Science Education*, 28 (491), pp. 520.
- Fearnley, A. (2000) *A comparability study in GCSE mathematics. A study based on the summer 1998 examination*. Organised by the Assessment and Qualification Alliance (Northern Examinations and Assessment Board) on behalf of the Joint Forum for the GCSE and GCE.
- Feilzer, M.Y. (2010) Doing mixed methods research pragmatically: Implications for the rediscovery of pragmatism as a research paradigm. *Journal of Mixed Methods Research*, 4 (6), pp. 6-16.
- Fisher-Hoch, H.; Hughes, S. and Bramley, T. (1997) What makes GCSE examination questions difficult? Outcomes of manipulating difficulty of GCSE questions. Paper presented at the British Educational Research Association Annual Conference, 11-14 September. University of York.
- Fowler, E.J. and Cannell, C.E. (1996) Using behavioral coding to identify cognitive problems with survey questions. In *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research*., eds. N. Schwarz and S. Sudman, San Francisco: Jossey-Bass, pp. 15-36.
- Fowles, D.E. (1995) A comparability study in Advanced level physics. A study based on the summer 1994 and 1990 examinations. Northern Examinations and Assessment Board on behalf of the Standing Research Advisory Committee of the GCE boards.
- Garnham, A. (1987) *Mental Models as Representations of Discourse and Text*. Chichester: Ellis Horwood.
- Gaskins, I.W., Guthrie, J.T., Satlow, E., Ostertag, J., Six, L., Byrne, J. and Connor, B. (1994) Integrating instruction of science, reading, and writing: Goals, teacher development, and assessment. *Journal of Research in Science Teaching*, 31, pp. 1039-1056.
- Gaskins, I.W., Guthrie, J.T., Satlow, E., Ostertag, J., Six, L., Byrne, J. and Connor, B. (1994) Integrating instruction of science, reading, and writing: Goals, teacher development, and assessment. *Journal of Research in Science Teaching*, 31, pp. 1039-1056.
- Gee, J.P. (2000) Discourse and sociocultural studies in reading . In *Handbook of reading research (Vol.3)*., eds. M.L. Kamil P. Mosenthal P.D. Pearson and R. Barr, Mahwah, NJ: Erlbaum, pp. 195-207.
- Georgas, J., Weiss, L.G., van de Vijver, F.J.R. and Saklofske, D.H. (eds) (2003) *Culture and Children's Intelligence: Cross-Cultural Analysis of the WISC-III*. San Diego, CA: Academic press.
- Ginsguger-Vogel, Y. and Astolfi, Y. (1987) Sur la lecture des manuels de biologie. *Alster*, 4, pp. 33-63.
- Goldstein, H. (1979) Consequences of Using the Rasch Model for Educational Measurement. *British Educational Research Journal*, 5 (2), pp. 211-220.
- Goldstein, H. (2004) International comparisons of student attainment: Some issues arising from the PISA study. *Assessment in Education: Principles, Policy and Practice*, 11, pp. 319-330.

- Goldstein, H., Bonnet, G. and Rocher, T. (2007) Multilevel Structural Equation Models for the Analysis of Comparative Data on Educational Performance. *Journal of Educational and Behavioral Statistics*, 32 (3), pp. 252-286.
- Gould, S.J. (1996) *The Mismeasure of Man*. 2nd ed. New York, NY: W.W. Norton.
- Graesser, A.C. and McNamara, D.S. (2011) Computational Analyses of Multilevel Discourse Comprehension. *Topics in Cognitive Science*, 3 (2), pp. 371-398.
- Graesser, A.C., Bommarreddy, S., Swamer, S. and Golding, J.M. (1996) Integrating questionnaire design with a cognitive computational model of human question answering. In *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research*, eds. N. Schwartz and S. Sudman, San Francisco: Jossey-Bass., pp. 143-174.
- Graesser, A.C., Cai, Z., Louwerse, M.M. and Daniel, F. (2006) Question Understanding Aid (QUAID). A web facility that tests question comprehensibility. *Public Opinion Quarterly*, 70 (1), pp. 3-22.
- Graesser, A.C., Kennedy, T., Wiemer-Hastings, P. and Ottati, V. (1999) The use of computational cognitive models to improve questions on surveys and questionnaires. In *Cognition and Survey Methods Research*, eds. M.G. Sirken D.J. Hermann S. Schechter N. Schwarz J.M. Tanur and R. Tourangeau, New York: Wiley, pp. 199-216.
- Graesser, A.C., Wiemer-Hastings, K., Kreuz, R., Wiemer-Hastings, P. and Marquis, K. (2000) QUAID: A questionnaire evaluation aid for survey methodologists. *Behavior Research Methods, Instruments, & Computers*, 32 (2), pp. 254-262.
- Graham, J.M. (2006) Congeneric and (Essentially) Tau-Equivalent Estimates of Score Reliability What They Are and How to Use Them. *Educational and Psychological Measurement*, 66 (6), pp. 930-944.
- Gray, E. (2000) *A comparability study in GCSE science 1998. A study based on the summer 1998 examination*. Organised by Oxford Cambridge and RSA Examinations (Midland Examining Group) on behalf of the Joint Forum for the GCSE and GCE.
- Greatorex, J., Elliott, G. and Bell, J.F. (2002) *A comparability study in GCE AS chemistry: A review of the examination requirements and a report on the cross-moderation exercise. A study based on the summer 2001 examination*. Organised by the Research and Evaluation Division, University of Cambridge Local Examinations Syndicate for Oxford Cambridge and RSA Examinations on behalf of the Joint Council for General Qualifications.
- Greatorex, J., Hamnett, L. and Bell, J.F. (2003) *A comparability study in GCE A level chemistry including the Scottish Advanced Higher grade. A review of the examination requirements and a report on the cross-moderation exercise. A study based on the summer 2002 examinations*. Organised by the Research and Evaluation Division, University of Cambridge Local Examinations Syndicate for Oxford Cambridge and RSA Examinations on behalf of the Joint Council for General Qualifications.
- Griffin, B., Carless, S. and Wilson, I. (2013) The effect of commercial coaching on selection test performance. *Medical Teacher*, 35, pp. 295-300.
- Griffiths, H.B. and McLone, R.R. (1979) *Qualities cultivated in mathematics degree examinations*. Southampton: University of Southampton.
- Grisay, A. and Monseur, C. (2007) Measuring the equivalence of item difficulty in the various versions of an international test. *Studies in Educational Evaluation*, 33, pp. 69-86.
- Grisay, A., de Jong, J.H.A.L., Gebhardt, E., Berezner, A. and Halleux-Monseur, B. (2007) Translation equivalence across PISA countries. *Journal of Applied Measurement*, 8 (3), pp. 249-266.
- Groves, F. (1995) Science vocabulary load of selected secondary science textbooks. *School Science and Mathematics*, 95, pp. 231-235.
- Groves, R.M. (1989) *Survey Errors and Survey Costs*. New York: Wiley.
- Hagquist, C. and Andrich, D. (2004) Is the Sense of Coherence-instrument applicable on adolescents? A latent trait analysis using Rasch-modelling. *Personality and Individual Differences*, 36 (4), pp. 955-968.
- Halliday, M.A.K. and Martin, J.R. (1993) *Writing Science: Literacy and Discursive Power*. Pittsburgh, PA: University of Pittsburgh.
- Hambleton, R.K. (1993) Translating achievement tests for use in cross-national studies. *European Journal of Psychological Assessment*, 9 (1), pp. 54-65.
- Hambleton, R.K. (2005) Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In *Adapting Educational and Psychological Tests for Cross-Cultural Assessment*, eds. R.K. Hambleton P.F. Merenda and C.D. Spielberger, Mahwah, NJ: Lawrence Erlbaum Associates, pp. 3-38.
- Hambleton, R.K. and Bollwark, J. (1991) Adapting tests for use in different cultures: Technical issues and methods. *Bulletin of the International Test Commission*, 18, pp. 3-32.

- Hambleton, R.K., Merenda, P.F. and Spielberger, C.D. (eds) (2005) *Adapting Educational and Psychological Tests for Cross-Cultural Assessment*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Hambleton, R.K. and Kanjee, A. (1995) Translating tests and attitude scales. In *International Encyclopedia of Education*, eds. T. Husen and T.N. Postlethwaite, 2nd ed. New York: Pergamon Press, pp. 6328-6334.
- Hanson, F.A. (2000) How tests create what they are intended to measure. In *Assessment: Social Practice and Social Product*, ed. Ann Filer, London: Routledge Falmer, pp. 67-81.
- Harkness, J. (ed.) (1998) *Cross-Cultural Equivalence*. Mannheim, Germany: ZUMA.
- Hauger, J.B. and Sireci, S.G. (2008) Detecting Differential Item Functioning Across Examinees Tested in Their Dominant Language and Examinees Tested in a Second Language. *International Journal of Testing*, 8 (3), pp. 237-250.
- Hermann-Abell, C.F. and DeBoer, G.E. (2011) Using distractor-driven standards-based multiple-choice assessments and Rasch modeling to investigate hierarchies of chemistry misconceptions and detect structural problems with individual items *Chemistry Education Research and Practice*, 12, pp. 184-192.
- Ho, D.Y.F. (1996) Filial piety and its psychological consequences. In *Handbook of Chinese Psychology*, ed. M.H. Bond, Hong Kong: Oxford University Press, pp. 155-165.
- Holbrook, J. and Rannikmae, M. (2007) The Nature of Science Education for Enhancing Scientific Literacy. *International Journal of Science Education*, 29 (11), pp. 1347-1362.
- Hölder, O. (1901) Die Axiome der Quantität und die Lehre vom Mass. *Berichte über die Verhandlungen der Königlich Sächsischen Gesellschaft der Wissenschaften zu Leipzig, Mathematisch-Physische Klasse*, , pp. 1.
- Holland, P.W. and Dorans, N.J. (2006) Linking and equating . In *Educational Measurement*, ed. R.L. Brennan, 4th ed. Westport, CT: American Council on Education. Praeger Press., pp. 187-220.
- Holland, P.W. and Thayer, D.T. (1988) Differential item performance and the Mantel-Haenszel procedure. In *Test Validity*, eds. H. Wainer and H. Braun, Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 129-145.
- Holland, P.W. and Wainer, H. (eds) (1993) *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Hopfenbeck, T.N. (2012) International datasets and comparisons in education. *Research Intelligence*, The role and value of international datasets and comparisons in education research (119), pp. 7-8.
- Hopfenbeck, T.N. and Maul, A. (2011) Examining Evidence for the Validity of PISA Learning Strategy Scales Based on Student Response Processes. *International Journal of Testing*, 11 (2), pp. 95-121.
- Horst, P. (1966) *Psychological Measurement and Prediction*. Belmont, California: Wadsworth Publishing Company.
- Houston, J.G. (1981) *Report of the inter-board cross-moderation study in economics at Advanced level*. Aldershot: Associated Examining Board.
- Howell, D. (1992) *Statistical Methods for Psychology*. 3rd ed. Belmont, CA: Duxbury Press.
- Huang, X., Wilson, M. and Wang, L. (2014) Exploring plausible causes of differential item functioning in the PISA science assessment: language, curriculum or culture. *Educational Psychology*, , pp. 1-13.
- Ibrahim, R. (2009) Psycholinguistic challenges in processing Arabic language. *International Journal of Psychology Research*, 4 (3), pp. 361-389.
- Impara, J.C. and Plake, B.S. (1998) Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement*, 35, pp. 69-81.
- International Test Commission (ITC) (2001) *International Test Commission Guidelines for Test Adaptation*. London: Author.
- Jones, B.E. (1993) *GCSE inter-group cross-moderation studies 1992. Summary report on studies undertaken on the summer 1992 examinations in English, mathematics and science*. Inter-Group Research Committee for the GCSE.
- Jones, B.E. (1997) *A review and evaluation of the methods used in the 1996 GCSE and GCE comparability studies*. Standing Committee on Research on behalf of the Joint Forum for the GCSE and GCE.
- Jones, B.P. (1988) Official Welsh. In *The Use of Welsh: A contribution to Sociolinguistics*, ed. M.J. Ball, Clevedon, : Multicultural Matters, pp. 172-181.
- Jones, D.V. (1998) National Curriculum tests for mathematics in English and Welsh: creating matched assessments. *Assessment in Education: Principles, Policy and Practice*, 5 (2), pp. 193-211.

- Kaufman, A.S. and Kaufman, N.L. (1990) *Kaufman Brief Intelligence Test—Manual*. Circle Pines, MN: American Guidance Service.
- Kelderman, H. (1989) Item bias detection using loglinear IRT. *Psychometrika*, 54, pp. 681-697.
- Kelley, T.L. (1927) *Interpretation of Educational Measurements*. Yonkers, NY: World Book Company.
- Kelly, G.A. (1955) *The psychology of personal constructs (Vols. I and II)*. New York: Norton.
- Kim, S. Kim, M. and Chun, M.M. (2005) Concurrent Working Memory Load Can Reduce Distraction. *Proceedings of the National Academy of Sciences of the United States of America*, 102 (45), pp. 16524-16529.
- Kirsch, I. (2001) *The International Adult Literacy Survey (IALS): Understanding What Was Measured. Research Report RR-01-25*. Princeton, NJ: Educational Testing Service.
- Kirsch, I., and Mosenthal, P.B. (1990) Exploring document literacy: Variables underlying the performance of young adults. *Reading Research Quarterly*, 25 (1), pp. 5-30.
- Kirsch, I., DeJong, J., Lafontaine, D., McQueen, J., Mendelovits, J. and Monseur, C. (2002) *Reading for change: Performance and engagement across countries: Results from PISA 2000*. Paris: Organisation for Economic Co-operation and Development.
- Kirsch, I., Jungeblut, A. and Mosenthal, P. (1998) The measurement of adult literacy. In *Adult Literacy in OECD countries.*, eds. T.S. Murray I.S. Kirsch and L.B. Jenkins, Washington, DC: National Center for Education Statistics.
- Kjærnsli, M. and Lie, S. (2004) PISA and scientific literacy: similarities and differences between the nordic countries. *Scandinavian Journal of Educational Research*, 48 (3), pp. 271-286.
- Kline, P. (1997) Commentary on Michell, Quantitative Science and the definition of measurement in psychology. *British Journal of Psychology*, 88, pp. 385-387.
- Kreiner, S. and 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), pp. 210-231.
- Kuhn, T. (1962) *The Structure of Scientific Revolutions*. Chicago, IL: University of Chicago Press.
- Kulik, J.A., Bangert-Drowns, R.L. and Kulik, C.C. (1984) Effectiveness of coaching for aptitude tests. *Psychological Bulletin*, 95 (2), pp. 179-188.
- Labasse, B. (1999) Perception et comprehension de l'ecrit. In *La Chose Imprimee*. Paris: Retz, pp. 458-462.
- LaBerge, D.S. and Samuels, S.J. (1974) Toward a theory of automatic information processing in reading. *Cognitive Psychology*, 6, pp. 293-323.
- Laming, D. (1997) A critique of a measurement-theoretic critique: Commentary on Michell, Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, 88, pp. 389-391.
- Landauer, T.K. (2012) *Improving text complexity measurement through the Reading Maturity Metric*. A paper presented at the annual meeting of the National Council on Measurement in Education; Vancouver, BC ed. http://images.pearsonassessments.com/images/tmrs/Word_Maturity_and_Text_Complexity_NCME.pdf: Pearson.
- Landauer, T.K., Kireyev, K. and Panaccione, C. (2011) Word Maturity: A New Metric for Word Knowledge. *Scientific Studies of Reading*, 15 (1), pp. 92-108.
- Lavie, N. (1995) Perceptual load as a necessary condition for selective attention. *Journal of Experimental Psychology: Human Perception and Performance*, 21, pp. 451-468.
- Lavie, N., Hirst, A., de Fockert, J.W. and Viding, E. (2004) Load theory of selective attention and cognitive control. *Journal of Experimental Psychology: General*, 133, pp. 339-254.
- Lawrence, I.M. (1993) *The Effect of Test Speededness on Subgroup Performance*. ETS Research Report 93-49 ed. Princeton, NJ: Educational Testing Service.
- Le, L.T. (2009) Investigating gender differential item functioning across countries and test languages for PISA science items. *International Journal of Testing*, 9 (2), pp. 122-133.
- Lederman, N.G. (1992) Students' and teachers' conceptions of the nature of science: A review of the research. *Journal of Research in Science Teaching*, 29 (4), pp. 331-359.
- Lehrer, A. (1990) Polysemy, conventionality, and the structure of the lexicon. *Cognitive Linguistics*, 1, pp. 207-246.
- Lenzner, T., Kaczmarek, L. and Lenzner, A. (2010) Cognitive burden of survey questions and response times: A psycholinguistic experiment. *Applied Cognitive Psychology*, 24 (7), pp. 1003-1020.
- Lessler, J.T. and Kalsbeek, W. (1993) *Nonsampling error in surveys*. New York: Wiley.
- Linacre, J.M. (1994) Sample size and item calibrations stability. *Rasch Measurement Transactions*, 7 (4), pp. 328.

- Linacre, J.M. (1998a) Detecting multidimensionality: Which residual data-types works best? . *Journal of Outcome Measurement*, 2, pp. 266-283.
- Linacre, J.M. (1998b) Structure in Rasch residuals: Why principal component analysis?. *Rasch Measurement Transactions*, 12, pp. 636.
- Linacre, J.M. (1999) Investigating rating scale category utility. *Journal of Outcome Measurement*, 3, pp. 103-122.
- Linacre, J.M. (2002) Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, 3 (1), pp. 85-106.
- Linn, R.L. (1993) Linking results of distinct assessments, 6, 83-102. *Applied Measurement in Education*, 6 (1), pp. 83-102.
- Lord, F.M. and Novick, M.R. (1968) *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Luce, R.D. (1997) Quantification and symmetry: Commentary on Michell, Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, 88, pp. 395-398.
- Luce, R.D. and Tukey, J.W. (1964) Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, 1 (1), pp. 1-27.
- Luce, R.D., Krantz, D.H., Suppes, P. and Tversky, A. (1990) *Foundations of Measurement*, vol. 3. . San Diego, CA: Academic Press.
- Lumley, T., Routitsky, A., Mendelovits, J. and Ramalingam, D. (2012) A framework for predicting item difficulty in reading tests. Paper presented at the American Educational Research Association (AERA) Meeting . Vancouver: .
- Luo, C.R. (1999) Semantic competition as the basis of Stroop interference: Evidence from color-word matching tasks. *Psychological Science*, 10, pp. 35-40.
- Luykx, A., Lee, O., Mahotiere, M., Lester, B., Hart, J. and Deaktor, R. (2007) Cultural and home language influences on children responses to science assessments. *Teachers College Record*, 109 (4), pp. 897-926.
- Malda, M., van de Vijver, F.J.R. and Temane, Q.M. (2010) Rugby versus Soccer in South Africa: Content familiarity contributes to cross-cultural differences in cognitive test scores. *Intelligence*, 38 (6), pp. 582-595.
- Mantel, N. and Haenszel, W. (1959) Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of National Cancer Institute*, 22, pp. 719-748.
- Marais, I. and Andrich, D. (2008a) Effects of varying magnitude and patterns of response dependence in the unidimensional Rasch model. *Journal of Applied Measurement*, 9 (2), pp. 1-20.
- Marais, I. and Andrich, D. (2008b) Formalising dimension and response violations of local independence in the unidimensional Rasch model. *Journal of Applied Measurement*, 9 (3), pp. 1-16.
- Martin, M.O., Gregory, K.D. and Stemler, S.E. (eds) (2000) *TIMSS 1999 Technical Report*. Chestnut Hill, MA: Boston College.
- Massey, A.J. (1979) *Comparing standards in English language: A report of the cross-moderation study based on the 1978 Ordinary level examinations of the nine GCE boards*. Bristol: Southern Universities' Joint Board and Test Development and Research Unit.
- McLone, R.R. and Patrick, H. (1990) *A study of the demands made by the two approaches to 'double Mathematics' . An investigation conducted by the University of Cambridge Local Examinations Syndicate on behalf of the Standing Research Advisory Committee of the GCE Examining Boards*. . Cambridge: University of Cambridge Local Examinations Syndicate.
- McQueen, J. and Mendolovits, J. (2003) PISA reading: cultural equivalence in a cross-cultural study. *Language Testing*, 20 (2), pp. 208-224.
- Mellenbergh, G.J. (1982) Contingency table models for assessing item bias. *Journal of Educational Statistics*, 7, pp. 105-118.
- Merzyn, G. (1987) The language of school science. *International Journal of Science Education*, 9, pp. 483-489.
- Meyer, L. (June 2003) *Repeat of AQA Standards Unit analyses originally run for the GCE Economics awarding meeting* . Internal unpublished AQA Standards Unit Paper ed.
- Michell, J. (1997a) Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, 88, pp. 355-383.
- Michell, J. (1997b) Reply to Kline, Laming, Lovie, Luce and Morgan. *British Journal of Psychology*, 88, pp. 401-406.

- Millar, R. (2006) Twenty First Century Science: Insights from the Design and Implementation of a Scientific Literacy Approach in School Science. *International Journal of Science Education*, 28 (13), pp. 1499-1521.
- Millar, R. and Osborne, J.F. (eds) (1998) *Beyond 2000: Science Education for the Future*. London: King's College London.
- Mislevy, R.J. (1991) Randomization-based inference about latent variables from complex samples. *Psychometrika*, 56, pp. 177-196.
- Mislevy, R.J. and Sheehan, K.M. (1987) Marginal estimation procedures, in A.E. Beaton (ed.). The NAEP 1983-84 technical report, National Assessment of Educational Progress, Princeton: Educational Testing Service, pp. 293-360.
- Mislevy, R.J., Beaton, A., Kaplan, B.A. and Sheehan, K. (1992) Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, 29 (2), pp. 133-161.
- Molenaar, I.W. (1995) Some background for item response theory and the Rasch model. In *Rasch models: Foundations, recent developments, and applications*, eds. G.H. Fischer and I.W. Molenaar, New York, NY: Springer, pp. 3-14.
- Morgan, M. (1997) Measurement in psychology: Commentary on Michell's Quantitative Science and the definition of measurement in psychology. *British Journal of Psychology*, 88, pp. 399-400.
- Moss, M.G. (2000) The language of school textbooks and the ideology of science. *Zona Proxima*, 1, pp. 45-55.
- Mullis, I.V.S. and Martin, M.O. (2000) Item analysis and review. In *TIMSS 1999. Technical Report*, eds. M.O. Martin D.G. Kelvin and S.E. Stemler, Chestnut Hill, MA: Boston College, pp. 225-230.
- Nardi, E. (2008) Cultural biases: a non-Anglophone perspective. *Assessment in Education: Principles, Policy & Practice*, 15 (3), pp. 259-266.
- National Research Council (NRC) (1996) *National Science Education Standards*. Washington, DC: National Academy of Sciences.
- Newton, P. (2005) Examination standards and the limits of linking. *Assessment in Education: Principles, Policy and Practice*, 12 (2), pp. 105-123.
- Newton, P. (2007) Clarifying the purposes of educational assessment. *Assessment in Education: Principles, Policy and Practice*, 14 (2), pp. 149-170.
- Newton, P. (2010) Thinking about linking. *Measurement: Interdisciplinary Research & Perspective*, 8 (1), pp. 38-56.
- Newton, P.E. and Shaw, S.D. (2014) *Validity in Educational and Psychological Assessment*. London, UK: Sage.
- Norman, O. (1998) Marginalized discourses and scientific literacy. *Journal of Research in Science Teaching*, 35, pp. 365-374.
- Norris, S.P. and Phillips, L.M. (2003) How literacy in its fundamental sense is central to scientific literacy. *Science Education*, 87 (2), pp. 224-240.
- Novak, J.D. (1977) *A Theory of Education*. London: Cornell University Press.
- Nygren, T.E. (1991) Psychometric properties of subjective workload measurement techniques: Implications for their use in the assessment of perceived mental workload. *Human Factors*, 33, pp. 17-33.
- Oakland, T. and Lane, H.B. (2004) Language, reading, and readability formulas: Implications for developing and adapting. *International Journal of Testing*, 4 (3), pp. 239-252.
- Oliveri, M.E. and Ercikan, K. (2011) Do different approaches to examining construct comparability in multilanguage assessments lead to similar conclusions? *Applied Measurement in Education*, 24 (4), pp. 349-366.
- Olson, D.R. (1996) Literate mentalities: Literacy, consciousness of language, and modes of thought. In *Modes of Thought*, eds. D.R. Olson and N. Torrance, Cambridge: Cambridge University Press, pp. 141-151.
- Organisation of Economic Co-operation and Development (1999) *Measuring Student Knowledge and Skills: A New Framework for Assessment*. Paris, France: Author.
- Organisation of Economic Co-operation and Development (2006) *Assessing Scientific, Reading and Mathematics Literacy: A Framework for PISA 2006*. Paris, France: Author.
- Organisation of Economic Co-operation and Development (2007a) *PISA 2006. Volume 2: Data/Données*. Paris, France: OECD.
- Organisation of Economic Co-operation and Development (2007b) *Science competencies for tomorrow's world. Volume 1: Analysis*. Paris, France: OECD.
- Organisation of Economic Co-operation and Development (2009a) *PISA 2006 Technical Report*. Paris, France: Organisation of Economic Co-operation and Development.

- Organisation of Economic Co-operation and Development (2009b) *PISA DATA ANALYSIS MANUAL*. 2nd ed. France: Paris: Author.
- Organisation of Economic Co-operation and Development (2010) *Viewing the United Kingdom School System through the Prism of PISA*. (www.oecd.org/pisa/46624007.pdf).
- Organisation of Economic Co-operation and Development (2013a) *PISA 2012 Assessment and Analytical Framework Mathematics, Reading, Science, Problem Solving and Financial Literacy*. Paris, France: OECD.
- Organisation of Economic Co-operation and Development (2013b) *Strong Performers and Successful Reformers in Education. Lessons from PISA for the United States*. Paris, France: OECD.
- Organisation of Economic Co-operation and Development (2014) *PISA 2012 Results in Focus. What 15-Year-Olds Know and What They Can Do With What They Know*. Paris, France: OECD.
- Paas, F. and van Merriënboer, J.J.G. (1993) The efficiency of instructional conditions: An approach to combine mental effort and performance measures. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 35, pp. 737-743.
- Paas, F. and van Merriënboer, J.J.G. (1994) Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Review*, 6 (51), pp. 71.
- Paas, F. and van Merriënboer, J.J.G. (1994) Variability of worked examples and transfer of geometrical problem solving skills: A cognitive-load approach. *Journal of Educational Psychology*, 86, pp. 122-133.
- Paas, F., Tuovinen, J.E., Tabbers, H. and Van Gerven, Pascal W. M. (2003) Cognitive Load Measurement as a Means to Advance Cognitive Load Theory. *Educational Psychologist*, 38 (1), pp. 63-71.
- Pallant, J.F. and Tennant, A. (2007) An introduction to the Rasch measurement model: An example using the Hospital Anxiety and Depression Scale (HADS). *British Journal of Clinical Psychology*, 46 (1), pp. 1-18.
- Panayides, P., Robinson, C. and Timms, P. (2010) The assessment revolution that has passed England by: Rasch measurement. *British Educational Research Journal*, 36 (4), pp. 611-626.
- Pinker, S. (1994) *The Language Instinct*. New York: William Morrow.
- Pinker, S. and Bloom, P. (1990) Natural language and natural selection. *Behavioral and Brain Sciences*, 13, pp. 707-726.
- Pollitt, A. and Ahmed, A. (1999) A new model of the question answering process. *International Association for Educational Assessment*, ed. University of Cambridge Local Examinations Syndicate, Bled, Slovenia.
- Pollitt, A., Ahmed, A. and Crisp, V. (2007) The demands on examination syllabuses and question papers. In *Techniques for Monitoring the Comparability of Examination Standards*, eds. P. Newton J. Baird H. Goldstein and H. Patrick, London: Qualifications and Curriculum Authority, pp. 166-206.
- Pollitt, A., and Ahmed, A. (2000) Comprehension failures in educational assessment. *Paper presented at the European Conference on Educational research*, Edinburgh: Research and Evaluation Division; University of Cambridge Local Examinations Syndicate, pp. 1.
- Pollitt, A., Entwistle, N.J., Hutchinson, C.J. and DeLuca, C. (1985) *What makes exam questions difficult?* Edinburgh: Scottish Academic Press.
- Pollitt, A., Hughes, A., Ahmed, A., Fisher-Hoch, H. and Bramley, T. (1998) *The effects of structure on the demands in GCSE and A level questions. Report to Qualifications and Curriculum Authority*.
- Poortinga, Y.H. (1989) Equivalence of cross-cultural data : An overview of basic issues. *International Journal of Psychology*, 24, pp. 737-756.
- Poortinga, Y.H. (1995) Cultural bias in assessment: Historical and thematic issues. *European Journal of Psychological Assessment*, 11 (3), pp. 140-146.
- Popper, K.R. (1959) *The Logic of Scientific Discovery*. London: Hutchinson Education.
- Powers, D.E. and Rock, D.A. (1999) Effects of coaching on SAT I: Reasoning test scores. *Journal of Educational Measurement*, 36 (2), pp. 93-118.
- Prais, S.J. (2003) Cautions on OECD's recent educational survey (PISA). *Oxford Review of Education*, 29, pp. 139-163.
- Rasch, G. (1960/1980) *Probabilistic Models for Some Intelligence and Attainment Tests (Copenhagen, Danish Institute for Educational Research.)* . Expanded edition (1980) foreword and afterword by B.D. Wright (1980) ed. Chicago: The University of Chicago Press.
- Rasch, G. (1961) On general laws and the meaning of measurement in psychology. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, ed. J. Neyman, Berkeley, CA: University of California Press., pp. 321.

- Ratcliffe, P. (1994) *A comparability study in GCSE geography. A study based on the summer 1993 examinations*. Northern Examinations and Assessment Board on behalf of the Inter-Group Research Committee for the GCSE.
- Raykov, T. (1997a) Estimation of composite reliability for congeneric measures . *Applied Psychological Measurement*, 21, pp. 173-184.
- Raykov, T. (1997b) Scale reliability, Cronbach's coefficient alpha, and violations of essential tau-equivalence with fixed congeneric components. *Multivariate Behavioral Research*, 32, pp. 329-353.
- Rayner, K., Pollatsek, A., Ashby, J. and Clifton, C.J. (2012) *Psychology of Reading*. 2nd ed. New York, NY: Psychology Press.
- Reckase, M.D. (1979) Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, 4, pp. 207-230.
- Rees, G., Frith, C.D. and Lavie, N. (1997) Modulating irrelevant motion perception by varying attentional load in an unrelated task. *Science*, 278, pp. 1616-1619.
- Rich, E.A. (1984) The gradual expansion of artificial intelligence. *IEEE Comput*, , pp. 4-12.
- Ross, V. (2014) Left in the past. *Scientific American Mind*, (July/August), pp. 10-11.
- Rupp, A.A., Garcia, P. and Jamieson, J. (2001) Combining multiple regression and CART to understand difficulty in second language reading and listening comprehension test items. *International Journal of Testing*, 1 (3 & 4), pp. 185-216.
- Sadler, P.M. (1998) Psychometric models for student-conceptions in science: Reconciling qualitative studies and distractor-driver assessment instruments. *Journal of Research in Science Teaching*, 35 (3), pp. 265-296.
- Sahlberg, P. (2011) PISA in Finland: An education miracle or an obstacle to change? *Center for Educational Policy Studies*, 1 (3), pp. 119-140.
- Sandilands, D., Oliveri, M.E., Zumbo, B.D. and Ercikan, K. (2013) Investigating Sources of Differential Item Functioning in International Large-Scale Assessments Using a Confirmatory Approach. *International Journal of Testing*, 13 (2), pp. 152-174.
- Schittekatte, M., Resing, W., Kort, W., Vermeir, G. and Verhaeghe, P. (2003) The Netherlands and Flemish-speaking Belgium. In *Culture and Children's Intelligence: Cross-Cultural Analysis of the WISC-III*, eds. J. Georgas L.G. Weiss F.J.R. van de Vijver and D.H. Saklofske, San Diego, CA: Academic Press, pp. 109-119.
- Schleicher, A. (2009) International Benchmarking as a Lever for Policy Reform . In *Change Wars.*, eds. A. Hargreaves and M. Fullan, Bloomington, IN: Solution Tree.
- Schober, M.E. and Conrad, E.G. (1997) Does conversational interviewing reduce survey measurement error? *Public Opinion Quarterly*, 60, pp. 576-602.
- Segall, M.H., Dasen, P.R., Berry, J.W. and Poortinga, Y.H. (1990) *Human Behavior in Global Perspective. An Introduction to Cross-Cultural Psychology*. New York, NY: Pergamon.
- Shiffrin, R.M. and Schneider, W. (1977) Controlled and automatic human information processing: II Perceptual learning, automatic attending, and a general theory. *Psychological Review*, 84, pp. 127-190.
- Sireci, S.G. (1997) Technical issues in linking tests across languages. *Educational Measurement: Issues and Practice*, 16 (1), pp. 12-19.
- Sireci, S.G. and Allalouf, A. (2003) Appraising item equivalence across multiple languages and cultures. *Language Testing*, 20 (2), pp. 148-166.
- Sireci, S.G., Patsula, L. and Hambleton, R.K. (2005) Statistical methods for identifying flaws in the test adaptation process. In *Adapting Educational and Psychological Tests for Cross-Cultural Assessment*, eds. R.K. Hambleton P.F. Merenda and C.D. Spielberger, Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 93-115.
- Smith, E. (2002) Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement*, 3 (2), pp. 205-231.
- Smith, R.M. and Hedges, L.V. (1982) Comparison of likelihood ratio χ^2 and Pearsonian χ^2 tests of fit in the Rasch model. *Education Research and Perspectives*, 9 (1), pp. 44-54.
- Solano-Flores, G. and Trumbull, E. (2003) Examining language in context: The need for new research and practice paradigms in the testing of English-Language Learners. *Educational Researcher*, 32 (2), pp. 3-13.
- Solano-Flores, G., Backhoff, E. and Contreras-Niño, L.Á. (2009) Theory of Test Translation Error. *International Journal of Testing*, 9 (2), pp. 78-91.
- Spielberger, C.D. (1972) *Anxiety: Current trends in theory and research: I*. New York: Academic Press.

- Stenner, J., Burdick, D.S. and Stone, M.H. (2008) Formative and reflective models: can a Rasch analysis tell the difference? *Rasch Measurement Transactions*, 22 (1), pp. 1152-1153.
- Stenner, J., Burdick, D.S. and Stone, M.H. (2009) Indexing vs. measuring. *Rasch Measurement Transactions*, 22 (4), pp. 1176-1177.
- Sternberg, R.J. (1977) *Intelligence, Information Processing, and Analogical Reasoning: The Componential Analysis of Human Abilities*. Hillsdale, NJ: Erlbaum.
- Sternberg, R.J. (1979) The nature of mental abilities. *American Psychologist*, 34, pp. 214-230.
- Steven, S.S. (1959) Measurement, psychophysics and utility. In *Measurement: Definitions and Theories.*, eds. C.W. Churchman and P. Ratoosh, New York, NY: Wiley, pp. 18-63.
- Stobart, G., Elwood, J., Jani, A. and Quinlan, M. (1994) *A comparability study in GCSE history: A study based on the summer 1993 examinations*. University of London Examinations and Assessment Council on behalf of the Inter-Group Research Committee for the GCSE.
- Stout, W. (2002) Psychometrics: from practice to theory and back. 15 Years of nonparametric multidimensional IRT, DIF/test equity, and skills diagnostic assessment. *Psychometrika*, 67, pp. 485-518.
- Stroop, J.R. (1935) Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 18 (6), pp. 643-662.
- Sutton, C. (1998) New perspectives on language in science. In *International Handbook of Science Education*, eds. B.J. Fraser and K.G. Tobin, Dordrecht: Kluwer, pp. 27-38.
- Swaffar, J.K. (1988) Readers, texts, and second languages: The interactive process. *The Modern Language Journal*, 2, pp. 123-149.
- Swaminathan, H. and Rogers, H.J. (1990) Detecting Differential Item Functioning Using Logistic Regression Procedures. *Journal of Educational Measurement*, 27 (4), pp. 361-370.
- Sweller, J., van Merriënboer, J.J.G. and Paas, F. (1998) Cognitive architecture and instructional design. *Educational Psychology Review*, 10, pp. 251-296.
- Takayama, K. (2008) The politics of international league tables: PISA in Japan's achievement crisis debate. *Comparative Education*, 44 (4), pp. 387-407.
- Tashakkori, A. and Teddlie, C. (eds) (2003) *Handbook of Mixed Methods in Social and Behavioral Research*. Thousand Oaks: Sage. Thousand Oaks, CA: Sage.
- Tennant, A. and Conaghan, P.G. (2007) The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper?. *Arthritis Rheum*, 57, pp. 1358-1362.
- Tennant, A. and Pallant, J.F. (2007) DIF matters: A practical approach to test if differential item functioning makes a difference. *Rasch Measurement Transactions*, 20 (4), pp. 1082-1084.
- Tewes, U. (2003) Germany. In *Culture and Children's Intelligence: Cross-Cultural Analysis of the WISC-III.*, eds. J. Georgas L.G. Weiss F.J.R. van de Vijver and D.H. Saklofske, San Diego, CA: Academic Press, pp. 129-135.
- Third International Mathematics and Science Study (1996a) *Mathematics Achievement in the Middle School Years: IEA's Third International Mathematics and Science Study (TIMSS)*. Chestnut Hill, MA: Center for the Study of Testing, Evaluation and Educational Policy, Boston College.
- Third International Mathematics and Science Study (1996b) *TIMSS Mathematics Items: Released Set for Population 2 (seventh and eighth grades).*. Chestnut Hill, MA: Center for the Study of Testing, Evaluation and Educational Policy, Boston College.
- Thomas, G. and Durant, J. (1987) Why should we promote the public understanding of science? *Scientific Literacy Papers*, 1, pp. 1-14.
- Traub, R.E. (1983) A priori considerations in choosing an Item Response Model. In *Applications of Item Response Theory.*, ed. R.K. Hambleton, Vancouver, BC.: Educational Research Institute of British Columbia, pp. 57-70.
- Traub, R.E. and Rowley, G.L. (1991) An NCME Instructional Module on Understanding Reliability. *Educational Measurement: Issues and Practice*, 10 (1), pp. 37-45.
- Traugott, E. (1978) On the expression of spatiotemporal relations in language. . In *Universals of Human Language: Word Structure.*, ed. J.H. Greenberg, Stanford, CA: Stanford Univ. Press., pp. 369-400.
- van de Vijver, F.J.R. and Leung, K. (1997) *Methods and Data Analysis for Cross- Cultural Research*. Newbury Park, CA: Sage.
- van de Vijver, F.J.R. and Poortinga, Y.H. (1997) Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment*, 13 (1), pp. 29-37.
- van de Vijver, F.J.R. and Tanzer, N.K. (2004) Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology*, 54, pp. 119-135.

- van de Vijver, F.J.R. and Poortinga, Y.H. (1991) Testing across Cultures. In *Advances in Educational and Psychological Testing.*, eds. R.K. Hambleton and J. Zaal, Boston, MA: Kluwer Academic, pp. 277-308.
- van de Vijver, F.J.R. and Poortinga, Y.H. (2005) Conceptual and methodological issues in adapting tests. In *Adapting Educational and Psychological Tests for Cross-Cultural Assessment.*, eds. R.K. Hambleton P.F. Merenda and C.D. Spielberger, Mahwah, NJ: Lawrence Erlbaum Associates, pp. 39-63.
- van Wyke, J. and Andrich, D. (2006) *A typology of polytomously scored mathematics items disclosed by the Rasch model: implications for constructing a continuum of achievement.* . http://www.mceecdya.edu.au/verve/_resources/ARC-Report02_A_Maths_typology.pdf.
- Velez, P. and Ashworth, S.D. (2007) The impact of item readability on the endorsement of the midpoint response in surveys . *Survey Research Methods*, 1, pp. 69-74.
- von Davier, A.A. (2010) What Dictates the Meaning of Test Linking? A Reaction to “Thinking About Linking”. *Measurement: Interdisciplinary Research & Perspective*, 8 (4), pp. 161-167.
- Walker, N.A., Forrest, G.M. and Kingdon, J.M. (1987) *Comparing the boards' examinations: An alternative approach.* Report on the second part (1984-1986) of an exercise conducted by the Standing Research Advisory Committee of the GCE boards involving the examinations in chemistry at Advanced level from 1978 to 1982 and from 1984 to 1986 inclusive.
- Waller, M.I. (1976) *Estimating Parameters in the Rasch model: Removing the Effects of Random Guessing.* Princeton, NJ: Educational Testing Service.
- Wang, X., Bradlow, E.T. and Wainer, H. (2002) A general Bayesian model for testlets: theory and applications. *Applied Psychological Measurement*, 26 (1), pp. 109-128.
- Welkowitz, J., Ewen, R.B. and Cohen, J. (1991) *Introductory Statistics for the Behavioral Sciences* . 4th ed. New York: Academic Press.
- Wendt, H., Bos, W. and Goy, M. (2011) On applications of Rasch models in international comparative large-scale assessments: a historical review. *Educational Research and Evaluation: An International Journal on Theory and Practice*, 17 (6), pp. 419-446.
- Westbury, I. (1992) Comparing American and Japanese Achievement: is the United States really a low achiever? . *Educational Researcher*, 21 (5), pp. 18-24.
- Whitburn, J. (1999) Why can't the English learn to subtract? In *Comparing Standards Internationally.*, eds. B. Jaworski and D. Phillips, Oxford: Symposium, pp. 163-182.
- Wickens, C.D. (2002) Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science*, 3 (2), pp. 159-177.
- Wickens, C.D. (2008) Human resources and mental workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 50, pp. 449-455.
- William, D. (2010) What counts as educational achievement? The role of constructs in the pursuit of equity in assessment. *Review of Research in Education*, 34, pp. 254-284.
- Wilson, M. and Adams, R.J. (1995) Rasch models for item bundles. *Psychometrika*, 60, pp. 181-198.
- Wober, M. (1974) Toward an understanding of the Kinganda concept of intelligence. In *Culture and Cognition: Reasoning in Cross-cultural Psychology.*, eds. J.W. Berry and P.R. Dasen, London: Methuen, pp. 261-280.
- Wolf, L.F., Smith, J.K. and Birnbaum, M.E. (1995) Consequence of performance, test, motivation, and mentally taxing items. *Applied Measurement in Education*, 8 (4), pp. 341-351.
- Wood, A. and Pollitt, A. (2006) Developing a methodology to enable students to participate in test development. *Paper presented at the Association for Educational Assessment - Europe Annual Conference* Naples, Italy: .
- Wood, T. and Wood, W. (1988) Assessing potential difficulties in comprehension fourth grade science textbooks. *Science Education*, 72, pp. 561-574.
- Wright, B.D. (1980) Afterword in *Probabilistic Models for Some Intelligence and Attainment Tests* by Rasch. Chicago: University of Chicago Press.
- Wright, B.D. (1997) A history of social science measurement. *Educational Measurement: Issues and Practice*, 16 (1), pp. 33-45.
- Wright, B.D. (1998) Where do dimensions come from? . *Popular Measurement*, XX, pp. 32.
- Wright, B.D. and Masters, G.N. (1982) *Rating Scale Analysis.* Rasch Measurement. Chicago, IL: MESA Press.
- Wright, B.D., Mead, R.H. and Draba, R. (1976) Detecting and correcting test item bias with a logistic response model: Research memorandum No 22. Chicago, IL: MESA Press.
- Wu, A.D. and Ercikan, K. (2006) Using Multiple-Variable Matching to Identify Cultural Sources of Differential Item Functioning. *International Journal of Testing*, 6 (3), pp. 287-300.

- Xie, Y. and Wilson, M. (2008) Investigating DIF and extensions using an LLTM approach and also an individual differences approach: An international testing context. *Psychology Science Quarterly*, 50(3): 403–416. *Psychology Science Quarterly*, 50 (3), pp. 403-416.
- Yore, L.D. and Treagust, D.F. (2006) Current Realities and Future Possibilities: Language and science literacy—empowering research and informing instruction. *International Journal of Science Education*, 28 (2-3), pp. 291-314.
- Zenisky, A.L., Hambleton, R.K. and Sireci, S.G. (2002) Identification and evaluation of local item dependencies in the medical college admissions test. *Journal of Educational Measurement*, 39(4), pp.291-309.

APPENDICES

Appendix 3.1

GREENHOUSE

Read the texts and answer the questions that follow.

THE GREENHOUSE EFFECT: FACT OR FICTION?

Living things need energy to survive. The energy that sustains life on the Earth comes from the Sun, which radiates energy into space because it is so hot. A tiny proportion of this energy reaches the Earth.

The Earth's atmosphere acts like a protective blanket over the surface of our planet, preventing the variations in temperature that would exist in an airless world.

Most of the radiated energy coming from the Sun passes through the Earth's atmosphere. The Earth absorbs some of this energy, and some is reflected back from the Earth's surface. Part of this reflected energy is absorbed by the atmosphere.

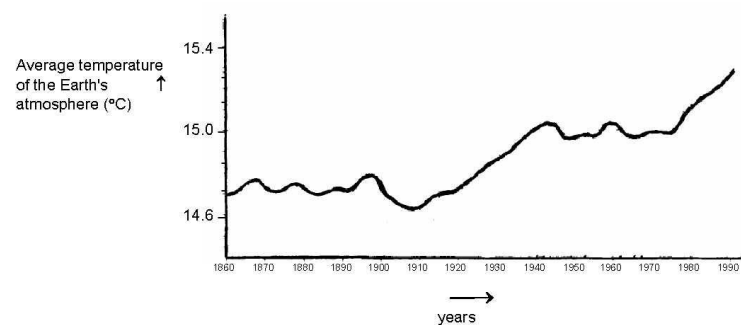
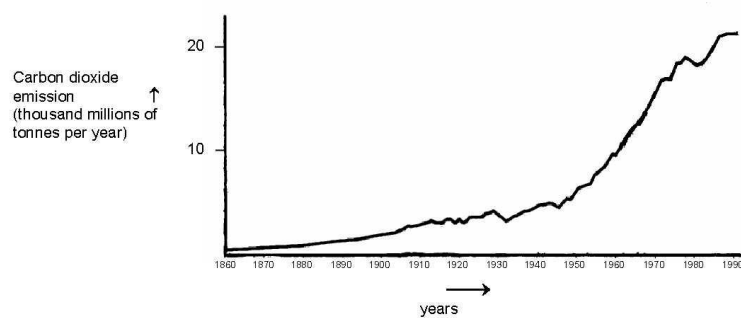
As a result of this the average temperature above the Earth's surface is higher than it would be if there were no atmosphere. The Earth's atmosphere has the same effect as a greenhouse, hence the term *greenhouse effect*.

The greenhouse effect is said to have become more pronounced during the twentieth century.

It is a fact that the average temperature of the Earth's atmosphere has increased. In newspapers and periodicals the increased carbon dioxide emission is often stated as the main source of the temperature rise in the twentieth century.

A student named André becomes interested in the possible relationship between the **average temperature of the Earth's atmosphere** and the **carbon dioxide emission** on the Earth.

In a library he comes across the following two graphs.



André concludes from these two graphs that it is certain that the increase in the **average temperature of the Earth's atmosphere** is due to the increase in the carbon dioxide emission.

Question 3: GREENHOUSE

S114Q03 – 01 02 11 12 99

What is it about the graphs that supports André's conclusion?

.....

.....

GREENHOUSE SCORING 3**Full Credit**

Code 11: Refers to the increase of both (average) temperature and carbon dioxide emission.

- As the emissions increased the temperature increased.
- Both graphs are increasing.
- Because in 1910 both the graphs began to increase.
- Temperature is rising as CO₂ is emitted.
- The information lines on the graphs rise together.
- Everything is increasing.
- The more CO₂ emission, the higher the temperature.

Code 12: Refers (in general terms) to a positive relationship between temperature and carbon dioxide emission.

[Note: This code is intended to capture students' use of terminology such as 'positive relationship', 'similar shape' or 'directly proportional'; although the following sample response is not strictly correct, it shows sufficient understanding to be given credit here.]

- The amount of CO₂ and average temperature of the Earth is directly proportional.
- They have a similar shape indicating a relationship.

No Credit

Code 01: Refers to the increase of either the (average) temperature or the carbon dioxide emission.

- The temperature has gone up.
- CO₂ is increasing.
- It shows the dramatic change in the temperatures.

Code 02: Refers to temperature and carbon dioxide emission without being clear about the nature of the relationship.

- The carbon dioxide emission (graph 1) has an effect on the earth's rising temperature (graph 2).
- The carbon dioxide is the main cause of the increase in the earth's temperature.

OR

Other responses.

- The carbon dioxide emission is greatly rising more than the average Earth's temperature. *[Note: This answer is incorrect because the extent to which the CO₂ emission and the temperature are rising is seen as the answer, rather than that they are both increasing.]*

- The rise of CO₂ over the years is due to the rise of the temperature of the Earth's atmosphere.
- The way the graph goes up.
- There is a rise.

Code 99: Missing.

Question 4: GREENHOUSE

S114Q04 – 0 1 2 9

Another student, Jeanne, disagrees with André's conclusion. She compares the two graphs and says that some parts of the graphs do not support his conclusion.

Give an example of a part of the graphs that does not support André's conclusion. Explain your answer.

.....

.....

.....

GREENHOUSE SCORING 4

Full Credit

- Code 2: Refers to one particular part of the graphs in which the curves are not both descending or both climbing and gives the corresponding explanation.
- In 1900–1910 (about) CO₂ was increasing, whilst the temperature was going down.
 - In 1980–1983 carbon dioxide went down and the temperature rose.
 - The temperature in the 1800's is much the same but the first graph keeps climbing.
 - Between 1950 and 1980 the temperature didn't increase but the CO₂ did.
 - From 1940 until 1975 the temperature stays about the same but the carbon dioxide emission shows a sharp rise.
 - In 1940 the temperature is a lot higher than in 1920 and they have similar carbon dioxide emissions.

Partial Credit

- Code 1: Mentions a correct period, without any explanation.
- 1930–1933.
 - before 1910.

Mentions only one particular year (not a period of time), with an acceptable explanation.

- In 1980 the emissions were down but the temperature still rose.

Gives an example that doesn't support André's conclusion but makes a mistake in mentioning the period. *[Note: There should be evidence of this mistake – e.g. an area clearly illustrating a correct answer is marked on the graph and then a mistake made in transferring this information to the text.]*

- Between 1950 and 1960 the temperature decreased and the carbon dioxide emission increased.

Refers to differences between the two curves, without mentioning a specific period.

- At some places the temperature rises even if the emission decreases.
- Earlier there was little emission but nevertheless high temperature.
- **When there is a steady increase in graph 1, there isn't an increase in graph 2, it stays constant.** *[Note: It stays constant "overall".]*
- Because at the start the temperature is still high where the carbon dioxide was very low.

Refers to an irregularity in one of the graphs.

- It is about 1910 when the temperature had dropped and went on for a certain period of time.
- In the second graph there is a decrease in temperature of the Earth's atmosphere just before 1910.

Indicates difference in the graphs, but explanation is poor.

- **In the 1940's the heat was very high** but the carbon dioxide very low. *[Note: The explanation is very poor, but the difference that is indicated is clear.]*

No Credit

Code 0: Refers to an irregularity in a curve without referring specifically to the two graphs.

- It goes a little up and down.
- It went down in 1930.

Refers to a poorly defined period or year without any explanation.

- The middle part.
- 1910.

Other responses.

- In 1940 the average temperature increased, but not the carbon dioxide emission.
- Around 1910 the temperature has increased but not the emission.

Code 9: Missing.

Question 5: GREENHOUSE

S114Q05 – 01 02 03 11 12 99

André persists in his conclusion that the **average temperature rise of the Earth's** atmosphere is caused by the increase in the carbon dioxide emission. But Jeanne **thinks that his conclusion is premature. She says: "Before accepting this conclusion you must be sure that other factors that could influence the greenhouse effect are constant".**

Name one of the factors that Jeanne means.

.....
.....

GREENHOUSE SCORING 5**Full Credit**

Code 11: Gives a factor referring to the energy/radiation coming from the Sun.

- The sun heating and maybe the earth changing position.
- Energy reflected back from Earth. *[Assuming that by "Earth" the student means "the ground".]*

Code 12: Gives a factor referring to a natural component or a potential pollutant.

- Water vapour in the air.
- Clouds.
- The things such as volcanic eruptions.
- Atmospheric pollution (gas, fuel).
- The amount of exhaust gas.
- **CFC's.**
- The number of cars.
- Ozone (as a component of air). *[Note: for references to depletion, use Code 03.]*

No Credit

Code 01: Refers to a cause that influences the carbon dioxide concentration.

- Clearing of rain forest.
- The amount of CO₂ being let off.
- Fossil fuels.

Code 02: Refers to a non-specific factor.

- Fertilisers.
- Sprays.
- How the weather has been.

Code 03: Other incorrect factors or other responses.

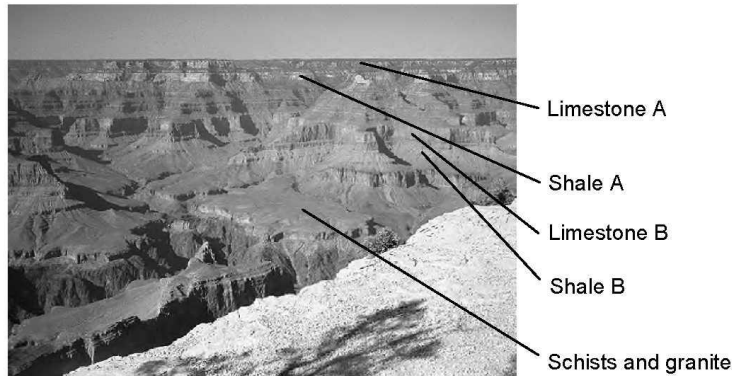
- Amount of oxygen.
- Nitrogen.
- The hole in the ozone layer is also getting bigger.

Code 99: Missing.

THE GRAND CANYON

The Grand Canyon is located in a desert in the USA. It is a very large and deep canyon containing many layers of rock. Sometime in the past, movements in the **Earth's crust lifted these layers up. The Grand Canyon is now 1.6 km deep in parts.** The Colorado River runs through the bottom of the canyon.

See the picture below of the Grand Canyon taken from its south rim. Several different layers of rock can be seen in the walls of the canyon.



Question 7: THE GRAND CANYON

S426Q07

About five million people visit the Grand Canyon national park every year. There is concern about the damage that is being caused to the park by so many visitors.

Can the following questions be answered by scientific investigation? Circle "Yes" or "No" for each question.

Can this question be answered by scientific investigation?	Yes or No?
How much erosion is caused by use of the walking tracks?	Yes / No
Is the park area as beautiful as it was 100 years ago?	Yes / No

THE GRAND CANYON SCORING 7

Full Credit

Code 1: Both correct: Yes, No in that order.

No Credit

Code 0: Other responses.

Code 9: Missing.

Question 3: THE GRAND CANYON

S426Q03

The temperature in the Grand Canyon ranges from below 0 °C to over 40 °C. Although it is a desert area, cracks in the rocks sometimes contain water. How do these temperature changes and the water in rock cracks help to speed up the breakdown of rocks?

- A Freezing water dissolves warm rocks.
- B Water cements rocks together.
- C Ice smoothes the surface of rocks.
- D Freezing water expands in the rock cracks.

THE GRAND CANYON SCORING 3**Full Credit**

Code 1: D. Freezing water expands in the rock cracks.

No Credit

Code 0: Other responses.

Code 9: Missing.

Question 5: THE GRAND CANYON

S426Q05

There are many fossils of marine animals, such as clams, fish and corals, in the Limestone A layer of the Grand Canyon. What happened millions of years ago that explains why such fossils are found there?

- A In ancient times, people brought seafood to the area from the ocean.
- B Oceans were once much rougher and sea life washed inland on giant waves.
- C An ocean covered this area at that time and then receded later.
- D Some sea animals once lived on land before migrating to the sea.

THE GRAND CANYON SCORING 5**Full Credit**

Code 1: C. An ocean covered this area at that time and then receded later.

No Credit

Code 0: Other responses.

Code 9: Missing.

Question 10S: THE GRAND CANYON

S426Q10S

How much do you agree with the following statements?

Tick only one box in each row.

	<i>Strongly Agree</i>	<i>Agree</i>	<i>Disagree</i>	<i>Strongly Disagree</i>
a) The systematic study of fossils is important.	☐ ₁	☐ ₂	☐ ₃	☐ ₄
b) Action to protect National Parks from damage should be based on scientific evidence.	☐ ₁	☐ ₂	☐ ₃	☐ ₄
c) Scientific investigation of geological layers is important.	☐ ₁	☐ ₂	☐ ₃	☐ ₄

Translation Note: In Question 10S, “National Parks” should be replaced with the most common term used in the country for nature or scenic reserves.

SUNSCREENS

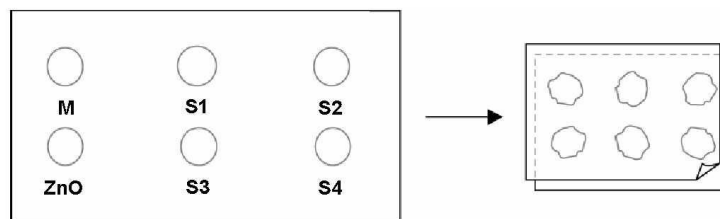
Mimi and Dean wondered which sunscreen product provides the best protection for their skin. Sunscreen products have a *Sun Protection Factor (SPF)* that shows how well each product absorbs the ultraviolet radiation component of sunlight. A high SPF sunscreen protects skin for longer than a low SPF sunscreen.

Mimi thought of a way to compare some different sunscreen products. She and Dean collected the following:

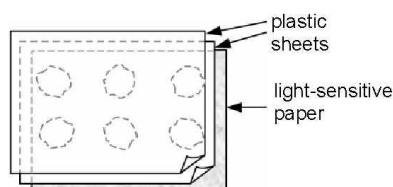
- two sheets of clear plastic that do not absorb sunlight;
- one sheet of light-sensitive paper;
- mineral oil (M) and a cream containing zinc oxide (ZnO); and
- four different sunscreens that they called S1, S2, S3, and S4.

Mimi and Dean included mineral oil because it lets most of the sunlight through, and zinc oxide because it almost completely blocks sunlight.

Dean placed a drop of each substance inside a circle marked on one sheet of plastic, then put the second plastic sheet over the top. He placed a large book on top of both sheets and pressed down.



Mimi then put the plastic sheets on top of the sheet of light-sensitive paper. Light-sensitive paper changes from dark grey to white (or very light grey), depending on how long it is exposed to sunlight. Finally, Dean placed the sheets in a sunny place.



Translation Note: If necessary, use the explicit translation “paper sensitive to light”, for “light-sensitive” paper. Do not use “photo-sensitive paper” as the translation.

Question 2: SUNSCREENS

S447Q02

Which one of these statements is a scientific description of the role of the mineral oil and the zinc oxide in comparing the effectiveness of the sunscreens?

- A Mineral oil and zinc oxide are both factors being tested.
- B Mineral oil is a factor being tested and zinc oxide is a reference substance.
- C Mineral oil is a reference substance and zinc oxide is a factor being tested.
- D Mineral oil and zinc oxide are both reference substances.

SUNSCREENS SCORING 2**Full Credit**

Code 1: D. Mineral oil and zinc oxide are both reference substances.

No Credit

Code 0: Other responses.

Code 9: Missing.

Question 3: SUNSCREENS

S447Q03

Which one of these questions were Mimi and Dean trying to answer?

- A How does the protection for each sunscreen compare with the others?
- B How do sunscreens protect your skin from ultraviolet radiation?
- C Is there any sunscreen that gives less protection than mineral oil?
- D Is there any sunscreen that gives more protection than zinc oxide?

SUNSCREENS SCORING 3**Full credit**

Code 1: A. How does the protection for each sunscreen compare with the others?

No credit

Code 0: Other responses.

Code 9: Missing.

Question 4: SUNSCREENS

S447Q04

Why was the second sheet of plastic pressed down?

- A To stop the drops from drying out.
- B To spread the drops out as far as possible.
- C To keep the drops inside the marked circles.
- D To make the drops the same thickness.

SUNSCREENS SCORING 4***Full Credit***

Code 1: D. To make the drops the same thickness.

No Credit

Code 0: Other responses.

Code 9: Missing.

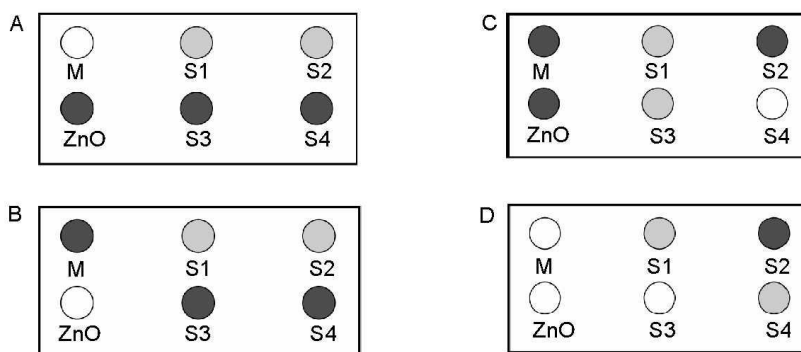
Question 5: SUNSCREENS

S447Q05 – 0 1 2 9

The light-sensitive paper is a dark grey and fades to a lighter grey when it is exposed to some sunlight, and to white when exposed to a lot of sunlight.

Translator's Note: The graphic used in the Field Trial MUST be replaced with the one below. Diagram C has been changed and the two grey shadings in all the diagrams have been modified so that they will be more distinct when printed.

Which one of these diagrams shows a pattern that might occur? Explain why you chose it.



Answer:

Explanation:

.....

.....

SUNSCREENS SCORING 5

Translator's Note: The layout of the scoring guide has been changed. It is no longer in table format.

Full Credit

- Code 2: A. With explanation that the ZnO spot has stayed dark grey (because it blocks sunlight) **and** the M spot has gone white (because mineral oil absorbs very little sunlight).
[It is not necessary (though it is sufficient) to include the further explanations that are shown in parentheses.]
- A. ZnO has blocked the sunlight as it should and M has let it through.
 - I chose A because the mineral oil needs to be the lightest shade while the zinc oxide is the darkest.

Partial Credit

Code 1: A. Gives a correct explanation for either the ZnO spot or the M spot, but **not both**.

- A. Mineral oil provides the lowest resistance against UVL. So with other substances the paper would not be white.
- A. Zinc oxide absorbs practically all rays and the diagram shows this.
- A because ZnO blocks the light and M absorbs it.

No Credit

Code 0: Other responses.

- A. *[No reason given.]*
- B. ZnO blocks the sunlight and mineral oil lets it through.

Code 9: Missing.

GENETICALLY MODIFIED CROPS

GM CORN SHOULD BE BANNED

Wildlife conservation groups are demanding that a new genetically modified (GM) corn be banned.

This GM corn is designed to be unaffected by a powerful new herbicide that kills conventional corn plants. This new herbicide will kill most of the weeds that grow in cornfields.

The conservationists say that because these weeds are feed for small animals, especially insects, the use of the new herbicide with the GM corn will be bad for the environment. Supporters of the use of the GM corn say that a scientific study has shown that this will not happen.

Here are details of the scientific study mentioned in the above article:

- Corn was planted in 200 fields across the country.
- Each field was divided into two. The genetically modified (GM) corn treated with the powerful new herbicide was grown in one half, and the conventional corn treated with a conventional herbicide was grown in the other half.
- The number of insects found in the GM corn, treated with the new herbicide, was about the same as the number of insects in the conventional corn, treated with the conventional herbicide.

Translator's Note: Corn is *Zea mays* and is referred to as maize in some countries.

Question 2: GENETICALLY MODIFIED CROPS

S508Q02

What factors were deliberately varied in the scientific study mentioned in the article?
Circle "Yes" or "No" for each of the following factors.

Was this factor deliberately varied in the study?	Yes or No?
The number of insects in the environment	Yes / No
The types of herbicide used	Yes / No

GENETICALLY MODIFIED CROPS SCORING 2

Full Credit

Code 1: Both correct: No, Yes in that order.

No Credit

Code 0: Other responses.

Code 9: Missing.

Question 3: GENETICALLY MODIFIED CROPS

S508Q03

Corn was planted in 200 fields across the country. Why did the scientists use more than one site?

- A So that many farmers could try the new GM corn.
- B To see how much GM corn they could grow.
- C To cover as much land as possible with the GM crop.
- D To include various growth conditions for corn.

GENETICALLY MODIFIED CROPS SCORING 3

Full Credit

Code 1: D. To include various growth conditions for corn.

No Credit

Code 0: Other responses.

Code 9: Missing.

Question 10N: GENETICALLY MODIFIED CROPS

S508Q10N

How much interest do you have in the following information?

Tick only one box in each row.

	<i>High Interest</i>	<i>Medium Interest</i>	<i>Low Interest</i>	<i>No Interest</i>
a) Learning about the process by which plants are genetically modified	☐ ₁	☐ ₂	☐ ₃	☐ ₄
b) Learning why some plants are not affected by herbicides	☐ ₁	☐ ₂	☐ ₃	☐ ₄
c) Understanding better the difference between cross-breeding and genetic modification of plants	☐ ₁	☐ ₂	☐ ₃	☐ ₄

L'EFFET DE SERRE

Lisez les textes suivants et répondez aux questions qui les accompagnent.

L'EFFET DE SERRE : RÉALITÉ OU FICTION ?

Les êtres vivants ont besoin d'énergie pour survivre. L'énergie qui alimente la vie sur Terre provient du Soleil, qui dégage de l'énergie dans l'espace, tant il est brûlant. Une infime proportion de cette énergie atteint la Terre.

L'atmosphère terrestre agit comme une couche de protection autour de la surface de la planète, empêchant les variations de température qui existeraient dans un monde sans air.

La plus grande partie de l'énergie venant du soleil traverse l'atmosphère terrestre. La Terre absorbe une partie de cette énergie, et une autre partie est réfléchiée et renvoyée par la surface de la Terre. Une partie de cette énergie réfléchiée par la Terre est absorbée par l'atmosphère.

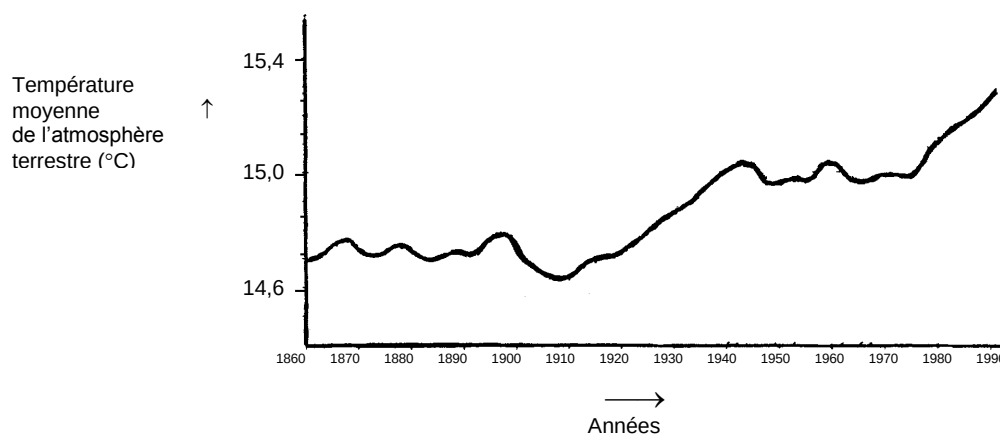
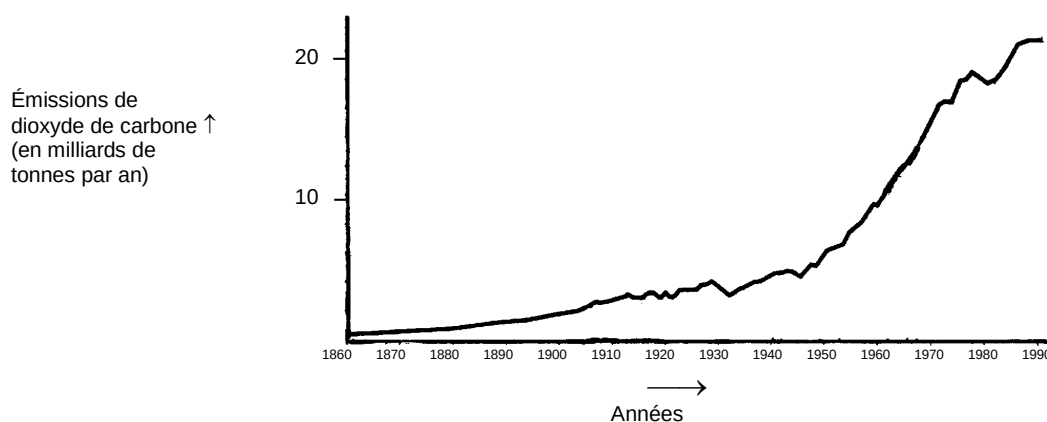
En conséquence, la température moyenne au-dessus de la surface terrestre est plus élevée qu'elle ne le serait s'il n'y avait pas d'atmosphère. L'atmosphère terrestre a le même effet qu'une serre, d'où l'expression «*effet de serre*».

L'effet de serre se serait intensifié au cours du vingtième siècle.

C'est un fait que la température moyenne de l'atmosphère de la Terre a augmenté. Les journaux et les magazines attribuent souvent à l'augmentation des émissions de dioxyde de carbone la principale responsabilité du réchauffement intervenu au vingtième siècle.

André, un étudiant, s'intéresse au rapport possible entre la température moyenne de l'atmosphère terrestre et l'émission de dioxyde de carbone sur Terre.

Dans une bibliothèque, il découvre les deux graphiques suivants.



André conclut, à partir de ces deux graphiques, qu'il est certain que la hausse de la température moyenne de l'atmosphère de la Terre est due à l'augmentation des émissions de dioxyde de carbone.

Question 3 : EFFET DE SERRE

S114Q03 - 01 02 11 12 99

Qu'est-ce qui, dans ces graphiques, confirme la conclusion d'André ?

.....
.....

EFFET DE SERRE : CONSIGNES DE CORRECTION Q 3

Crédit complet

Code 11 : Fait référence au fait que, tant la température (moyenne) que les émissions de dioxyde de carbone ont augmenté.

- Quand les émissions se sont accrues, la température est montée.
- Les deux graphiques montrent une augmentation.
- Parce qu'en 1910 les deux courbes ont commencé à croître.
- La température monte quand il y a des émissions de CO₂.
- Les courbes sur les graphiques montent en même temps.
- Tout augmente.
- Plus il y a d'émissions de CO₂, plus la température est élevée.

Code 12 : Fait référence (en termes généraux) à un rapport positif entre la température et les émissions de dioxyde de carbone.

[Note : ce code a pour but de relever l'usage que les élèves font de termes comme « relation positive », « forme semblable » ou « directement proportionnel » : bien que l'exemple de réponse suivant ne soit pas à proprement parler correct, il manifeste une compréhension suffisante pour se voir attribuer un crédit complet dans ce cas précis].

- La quantité de CO₂, et la température moyenne de la Terre sont directement proportionnelles
- Elles suivent un tracé analogue qui indique un rapport entre elles.

Pas de crédit

Code 01 : Fait référence à l'augmentation soit de la température (moyenne), soit des émissions de dioxyde de carbone.

- La température a monté.
- Le CO₂ augmente.
- Cela indique un changement radical de la température.

Code 02 : Fait référence à la température et à l'émission de dioxyde de carbone sans spécifier clairement la nature de la relation.

- L'émission de dioxyde de carbone (diagramme 1) a un effet sur l'augmentation de la température terrestre (diagramme 2).
- Le dioxyde de carbone est la cause principale de l'augmentation de la température sur la Terre.

OU

Autres réponses.

- Les émissions de dioxyde de carbone augmentent de façon spectaculaire, plus encore que la température de la Terre *[Note : cette réponse est incorrecte parce que la mesure dans laquelle les émissions de CO₂ et la température de la Terre augmentent est perçue comme étant la réponse, et non le fait qu'elles augmentent l'une et l'autre].*
- L'augmentation de CO₂ au fil des ans est due à une hausse de température de l'atmosphère terrestre.

- La façon dont le graphique grimpe.
- Il y a une hausse.

Code 99 : Omission.

Question 4 : EFFET DE SERRE

S114Q04 - 02 03 11 12 13 14 15 21 99

Jeanne, une autre élève, n'est pas d'accord avec la conclusion d'André. Elle compare les deux graphiques et dit que certaines parties de ceux-ci ne confirment pas sa conclusion.

Donnez un exemple, en citant une partie de ces graphiques qui ne confirme pas la conclusion d'André. Expliquez votre réponse.

.....

.....

.....

EFFET DE SERRE : CONSIGNES DE CORRECTION Q 4

Crédit complet

Code 21 : Fait référence à une partie spécifique du graphique dans laquelle les courbes ne sont pas toutes deux ascendantes ou descendantes, et fournit une explication en rapport avec le phénomène constaté.

- En 1900–1910 (environ) le CO₂ a augmenté, alors que la température a continué à descendre.
- En 1980–1983, le dioxyde de carbone a diminué tandis que la température a augmenté.
- Pendant les années 1800, la température reste assez stable mais la courbe du premier graphique est continuellement ascendante.
- Entre 1950 et 1980, la température n'a pas augmenté alors que le CO₂ a augmenté.
- La température est plus ou moins constante de 1940 à 1975, tandis que les émissions de dioxyde de carbone sont en forte augmentation.
- De 1860 à 1900, la courbe du CO₂ est continue, alors que celle de la température est très irrégulière.
- En 1940, la température est beaucoup plus élevée qu'en 1920, tandis que les émissions de dioxyde de carbone sont similaires.

Crédit partiel

Code 11 : Cite une période correcte sans fournir d'explication.

- 1930 - 1933.
- Aux alentours de 1910.

Code 12 : Ne mentionne qu'une année particulière (pas une période), avec une justification acceptable

[Note : Utilisez le code 14 si l'explication fait référence à une irrégularité dans un des graphiques].

- En 1980, le niveau d'émissions a été bas, mais la température a continué à monter.
- En 1910, le dioxyde de carbone a augmenté alors que la température était en baisse.

OU

Donne un exemple qui ne confirme pas la conclusion d'André, mais fait une erreur en citant la période.

- Entre 1950 et 1960, la température a baissé et les émissions de dioxyde de carbone ont augmenté.

Code 13 : Fait référence à la différence entre les deux courbes, sans mentionner de période spécifique.

- À certains moments, la température est en hausse même quand les émissions sont en baisse.
- Auparavant, il y avait peu d'émissions et pourtant la température était élevée.
- Ils n'augmentent pas au même rythme.
- Tandis que le graphique 1 montre une hausse constante, il n'y a pas de véritable hausse dans le graphique 2, qui reste constant [Note : *il reste constant «dans l'ensemble»*].
- Parce qu'au début, la température est encore assez élevée alors qu'il y avait très peu de dioxyde de carbone.

Code 14 : Fait référence à une irrégularité dans un des graphiques.

- C'est à peu près en 1910 que la température a chuté et cela a duré un certain temps.
- Dans le second graphique, il y a une baisse de la température de l'atmosphère terrestre vers 1910.

Code 15 : Indique une divergence entre les graphiques, mais l'explication est très faible.

- Pendant les années quarante, la chaleur était très élevée, mais le taux de dioxyde de carbone était très bas [Note : *l'explication est faible, mais la différence citée est claire*].

Pas de crédit

Code 01 : Fait référence à une irrégularité dans une des courbes sans faire spécifiquement référence aux deux graphiques.

- Cela monte et descend un peu.
- C'est descendu en 1930.

Code 02 : Fait référence à une période mal définie ou à une année sans fournir d'explication.

- La partie centrale.
- 1910.

Code 03 : Autres réponses.

- En 1940, la température moyenne a augmenté, mais pas les émissions de dioxyde de carbone.
- Autour de 1910 la température est montée mais pas les émissions.

Code 99 : Omission.

Question 5 : EFFET DE SERRE

S114Q05 - 01 02 03 11 12 99

André maintient sa conclusion : le réchauffement de l'atmosphère est dû à l'augmentation des émissions de dioxyde de carbone. Mais Jeanne pense que sa conclusion est prématurée. Elle dit : « *Avant d'accepter cette conclusion, tu dois t'assurer que d'autres facteurs qui pourraient avoir une influence sur l'effet de serre sont constants* ».

Citez un des facteurs auxquels Jeanne fait allusion.

.....

.....

EFFET DE SERRE : CONSIGNES DE CORRECTION Q 5

Crédit complet

Code 11 : Cite un facteur qui fait référence à l'énergie/au rayonnement solaire :

- La chaleur du soleil et peut-être un changement de position de la terre.
- L'énergie solaire réfléchie par la Terre.

Code 12 : Cite un facteur qui fait référence à une composante naturelle ou à un agent polluant potentiel.

- La vapeur d'eau dans l'air.
- Les nuages.
- Les phénomènes comme les éruptions volcaniques.
- La pollution atmosphérique (gaz, pétrole).
- La quantité de gaz d'échappement.
- Les CFC.
- Le nombre d'automobiles.
- L'ozone (en tant que composante de l'air) [*Note : utilisez le code 03 si la réponse fait référence à la diminution de la couche d'ozone*].

Pas de crédit

Code 01 : Fait référence à une cause qui a une influence sur la concentration de dioxyde de carbone.

- La destruction de la forêt tropicale.
- La quantité de CO₂ dégagée.
- Les combustibles fossiles.

Code 02 : Fait référence à un facteur non spécifique.

- Les engrais.
- Les aérosols.
- Les conditions météo.

Code 03 : Autres facteurs incorrects ou autres réponses.

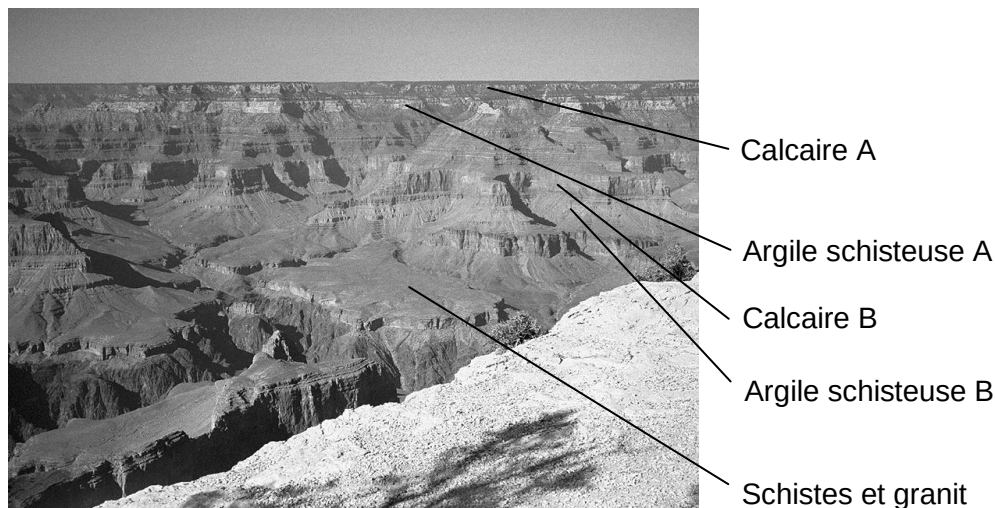
- La quantité d'oxygène.
- L'azote.
- Le trou dans la couche d'ozone est aussi en train de devenir plus grand.

Code 99 : Omission.

LE GRAND CANYON

Le Grand Canyon est situé dans un désert des États-Unis d'Amérique. C'est un canyon très vaste et très profond, constitué de nombreuses couches rocheuses. Autrefois, des mouvements de la croûte terrestre ont soulevé ces couches. Le Grand Canyon atteint à présent jusqu'à 1,6 km de profondeur à certains endroits. Le fleuve Colorado coule au fond du canyon.

La photo du Canyon que vous voyez ci-dessous a été prise du versant sud. On distingue différentes couches rocheuses formant les parois du canyon.



Question 1 : GRAND CANYON

S426Q01 – 0 1 9

À cause de quoi le Grand Canyon s'est-il formé ?

.....

.....

GRAND CANYON : CONSIGNES DE CORRECTION Q 1

Crédit complet

Code 1 : Le fleuve Colorado a creusé les couches successives en érodant la roche.

- Le fleuve a creusé son lit à travers les couches rocheuses.
- L'érosion par l'eau, parce qu'il y a un fleuve au fond.
- Le mouvement de la croûte terrestre a soulevé ces couches ; et l'érosion.
- Le fleuve a modelé le terrain.
- L'eau qui coule.

Note de correction : Pour obtenir le code 1, la réponse doit faire référence à l'érosion par l'eau (ou simplement à « l'érosion »), ou mentionner explicitement l'*action* de l'eau.

Pas de crédit

Code 0 : Autres réponses.

- Des canyons souterrains qui se sont écroulés.
- L'érosion par le vent.

- À cause de l'eau.
- Le fleuve.
- Le Grand Canyon s'est formé parce que le mouvement de la croûte terrestre a soulevé les couches rocheuses qui composent les parois du canyon.

Code 9 : Omission.

Question 7 : GRAND CANYON

S426Q07

Environ cinq millions de personnes visitent le parc national du Grand Canyon chaque année. On s'inquiète des dégâts qui sont causés au parc par tant de visiteurs.

Peut-on répondre aux questions suivantes grâce à une étude scientifique ? Entourez « Oui » ou « Non » pour chacune des questions.

Peut-on répondre à cette question grâce à une étude scientifique ?	Oui ou Non?
Quelle est l'ampleur de l'érosion causée par l'utilisation des sentiers de promenade ?	Oui / Non
Est-ce que le parc est aussi beau aujourd'hui qu'il y a 100 ans ?	Oui / Non

GRAND CANYON : CONSIGNES DE CORRECTION Q 7

Crédit complet

Code 1 : Les deux réponses sont correctes. Dans l'ordre : Oui, Non.

Pas de crédit

Code 0 : Autres réponses.

Code 9 : Omission.

Question 3 : GRAND CANYON

S426Q03

La température dans le Grand Canyon varie de moins de 0 °C à plus de 40 °C. Bien que la zone soit désertique, les fissures de la roche contiennent parfois de l'eau. De quelle façon ces changements de température et l'eau dans les fissures contribuent-elles à accélérer l'effritement de la roche ?

- A En gelant, l'eau dissout les roches chaudes.
- B L'eau cimente les roches entre elles.
- C La glace polit la surface des roches.
- D En gelant, l'eau se dilate dans les fissures de la roche.

GRAND CANYON : CONSIGNES DE CORRECTION Q 3

Crédit complet

Code 1 : D. En gelant, l'eau se dilate dans les fissures de la roche.

Pas de crédit

Code 0 : Autres réponses.

Code 9 : Omission.

Question 5 : GRAND CANYON

S426Q05

Dans la couche calcaire A du Grand Canyon, se trouvent de nombreux fossiles d'animaux marins, comme des palourdes, des poissons et des coraux. Que s'est-il passé il y a des millions d'années pour que ces fossiles se trouvent là ?

- A D'anciennes peuplades ont transporté des produits de la mer depuis l'océan jusqu'à cette région.
- B Autrefois, les océans étaient beaucoup plus agités : des vagues géantes emportaient des animaux marins jusqu'à l'intérieur des terres.
- C À cette époque, un océan recouvrait la région et, plus tard, il s'est retiré.
- D Certains animaux marins ont vécu sur terre avant de migrer vers les mers.

GRAND CANYON : CONSIGNES DE CORRECTION Q 5

Crédit complet

Code 1 : C. À cette époque, un océan recouvrait la région et, plus tard, il s'est retiré.

Pas de crédit

Code 0 : Autres réponses.

Code 9 : Omission.

Question 10S : LE GRAND CANYON

S426Q10S

Dans quelle mesure êtes-vous d'accord avec les affirmations suivantes ?

Ne cochez qu'une case par ligne.

	<i>Tout à fait d'accord</i>	<i>D'accord</i>	<i>Pas d'accord</i>	<i>Pas du tout d'accord</i>
a) L'étude systématique des fossiles est importante.	☐ ₁	☐ ₂	☐ ₃	☐ ₄
b) Les mesures de protection des parcs naturels contre les dégâts doivent s'appuyer sur des preuves scientifiques.	☐ ₁	☐ ₂	☐ ₃	☐ ₄
c) L'étude scientifique des couches géologiques est importante.	☐ ₁	☐ ₂	☐ ₃	☐ ₄

Note de traduction : Dans la Question 10S, le terme « Parcs naturels » (“National Parks” dans la version anglaise) doit être traduit par un terme couramment utilisé dans votre pays pour désigner les parcs naturels.

ÉCRANS SOLAIRES

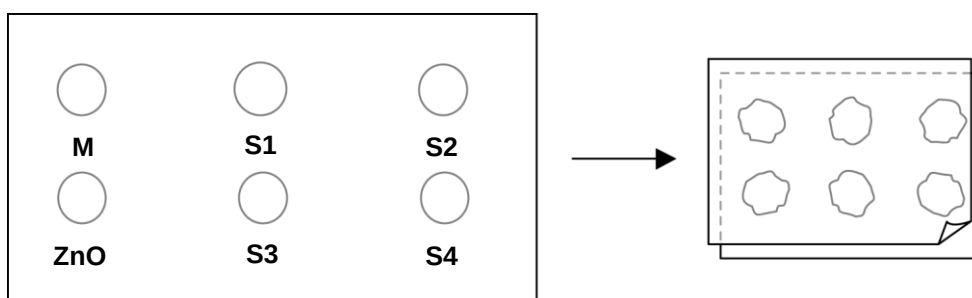
Mimi et David se demandent quel écran solaire offre la meilleure protection à leur peau. Les écrans solaires ont un *facteur de protection solaire (FPS)* indiquant dans quelle mesure ils absorbent les rayons ultraviolets de la lumière du soleil. Un écran solaire à FPS élevé protège la peau plus longtemps qu'un écran solaire à faible FPS.

Mimi a imaginé une manière de comparer divers écrans solaires. David et elle ont rassemblé le matériel suivant :

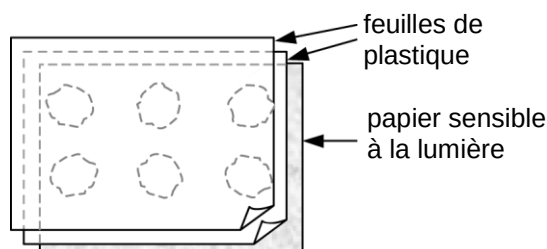
- deux feuilles de plastique transparent qui n'absorbent pas la lumière du soleil ;
- une feuille de papier sensible à la lumière ;
- de l'huile minérale (M) et une crème contenant de l'oxyde de zinc (ZnO) ;
- quatre écrans solaires différents qu'ils ont nommés S1, S2, S3, et S4.

Mimi et David ont utilisé l'huile minérale parce qu'elle laisse passer presque toute la lumière du soleil et l'oxyde de zinc parce qu'il bloque presque complètement la lumière du soleil.

David a déposé une goutte de chaque substance dans un des cercles tracés sur une des feuilles de plastique, qu'il a ensuite recouverte avec la seconde feuille de plastique. Il a placé un grand livre sur les deux feuilles et a appuyé dessus.



Ensuite, Mimi a posé les feuilles de plastique sur le papier sensible à la lumière. Le papier sensible à la lumière a la propriété de passer du gris foncé au blanc (ou au gris très clair) en fonction de la durée de son exposition à la lumière du soleil. Enfin, David a placé les feuilles dans un endroit ensoleillé.



Note de traduction. Si cela s'avère nécessaire, traduisez l'expression anglaise « light-sensitive paper » de manière explicite comme en français (papier sensible à la lumière). Veuillez ne pas traduire cette expression par « papier photosensible ».

Question 2 : ÉCRANS SOLAIRES

S447Q02

Parmi les énoncés suivants, lequel est une description scientifique du rôle de l'huile minérale et de l'oxyde de zinc dans la comparaison de l'efficacité des écrans solaires ?

- E L'huile minérale et l'oxyde de zinc sont tous deux des facteurs que l'on teste.
- F L'huile minérale est un facteur que l'on teste et l'oxyde de zinc est une substance de référence.
- G L'huile minérale est une substance de référence et l'oxyde de zinc est un facteur que l'on teste.
- H L'huile minérale et l'oxyde de zinc sont tous deux des substances de référence.

ÉCRANS SOLAIRES : CONSIGNES DE CORRECTION Q 2***Crédit complet***

Code 1 : D. L'huile minérale et l'oxyde de zinc sont tous deux des substances de référence.

Pas de crédit

Code 0 : Autres réponses.

Code 9 : Omission.

Question 3 : ÉCRANS SOLAIRES

S447Q03

À laquelle des questions suivantes Mimi et David ont-ils essayé de répondre ?

- I Quelle protection chaque écran solaire offre-t-il par comparaison avec les autres ?
- J Comment les écrans solaires protègent-ils la peau contre les rayons ultraviolets ?
- K Parmi les écrans solaires, y en a-t-il un qui protège moins que l'huile minérale ?
- L Parmi les écrans solaires, y en a-t-il un qui protège davantage que l'oxyde de zinc ?

ÉCRANS SOLAIRES : CONSIGNES DE CORRECTION Q 3***Crédit complet***

Code 1 : A. Quelle protection chaque écran solaire offre-t-il par comparaison avec les autres ?

Pas de crédit

Code 0 : Autres réponses.

Code 9 : Omission.

Question 4 : ÉCRANS SOLAIRES**S447Q04**

Pourquoi a-t-on appuyé sur la seconde feuille de plastique ?

- M Pour empêcher les gouttes de sécher.
- N Pour étaler les gouttes le plus possible.
- O Pour maintenir les gouttes à l'intérieur des cercles tracés.
- P Pour donner aux gouttes la même épaisseur.

ÉCRANS SOLAIRES : CONSIGNES DE CORRECTION Q 4***Credit complet***

Code 1 : D. Pour donner aux gouttes la même épaisseur.

Pas de crédit

Code 0 : Autres réponses.

Code 9 : Omission.

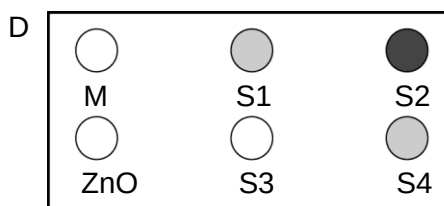
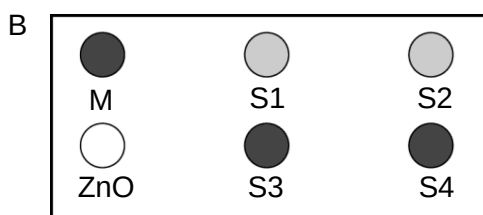
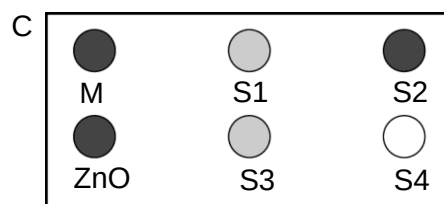
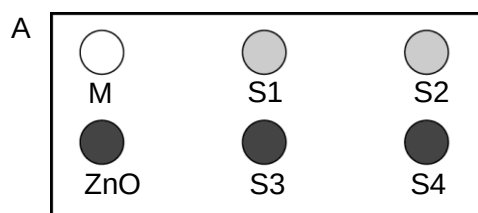
Question 5 : ÉCRANS SOLAIRES

S447Q05 – 0 1 2 9

Le papier sensible à la lumière est gris foncé ; il devient gris clair quand il est exposé à un peu de lumière du soleil et blanc quand il est exposé à beaucoup de lumière de soleil.

Note de traduction : Il FAUT remplacer les schémas utilisés lors de l'Essai de Terrain par ceux ci-dessous. Le schéma C a été changé et les deux nuances de gris utilisées dans chacun des schémas ont été modifiées afin qu'elles se distinguent mieux l'une de l'autre à l'impression.

Parmi ces schémas, lequel présente les résultats que l'on pourrait obtenir ? Expliquez pourquoi vous l'avez choisi.



Réponse :

Explication :

.....

.....

ÉCRANS SOLAIRES : CONSIGNES DE CORRECTION Q 5

Note de traduction : La mise en page des consignes de correction a été modifiée. Elle n'apparaît plus sous forme de tableau.

Crédit complet

Code 2 : A. Explique que le cercle ZnO est resté gris foncé (parce qu'il bloque la lumière du soleil) **et** que le cercle M est devenu blanc (parce que l'huile minérale absorbe très peu de lumière du soleil).

*[Il n'est **pas** nécessaire (quoique suffisant) que les explications complémentaires entre parenthèses soient fournies.]*

- A. Le ZnO a bloqué la lumière du soleil comme prévu et M l'a laissé passer.
- J'ai choisi A parce que l'huile minérale doit être la plus claire alors que l'oxyde de zinc est le plus foncé.

Crédit partiel

Code 1 : A. Donne une explication correcte soit pour le cercle ZnO, **soit** pour le cercle M, mais **pas** pour les deux, **et** ne donne pas d'explication incorrecte pour l'autre cercle.

- A. L'huile minérale a la plus faible résistance aux rayons UV. Donc, le papier ne serait pas blanc pour les autres substances.
- A. L'oxyde de zinc absorbe presque tous les rayons, et le diagramme le montre.

Pas de crédit

Code 0 : Autres réponses.

- A. Parce que le ZnO bloque la lumière et que M l'absorbe.
- B. Le ZnO bloque la lumière du soleil et l'huile minérale la laisse passer.

Code 9 : Omission.

CULTURES GENETIQUEMENT MODIFIÉES

LE MAÏS OGM DEVRAIT ÊTRE INTERDIT

Des groupes de protection de la nature ont demandé l'interdiction d'une nouvelle espèce de maïs génétiquement modifiée (OGM, organisme génétiquement modifié).

Ce maïs OGM est conçu pour résister à un nouvel herbicide puissant qui détruit les plants de maïs traditionnels. Ce nouvel herbicide détruira la plupart des mauvaises herbes qui poussent dans les champs de maïs.

Les protecteurs de la nature déclarent que, comme ces mauvaises herbes sont une source de nourriture pour les petits animaux, en particulier les insectes, l'utilisation de ce nouvel herbicide avec le maïs OGM nuira à l'environnement. Les partisans du maïs OGM répondent qu'une étude scientifique a démontré que cela n'arrivera pas.

Voici quelques détails de l'étude scientifique mentionnée dans l'article ci-dessus :

- On a semé du maïs dans 200 champs à travers le pays.
 - On a divisé chaque champ en deux parties. Dans une moitié, on a cultivé du maïs génétiquement modifié (OGM) traité avec le nouvel herbicide puissant, et dans l'autre moitié on a cultivé du maïs traditionnel traité avec un herbicide traditionnel.
 - On a trouvé à peu près le même nombre d'insectes sur le maïs OGM traité avec le nouvel herbicide que sur le maïs traditionnel traité avec l'herbicide traditionnel.

Note de traduction : le nom scientifique du maïs est Zea mays.

Question 2 : CULTURES GÉNÉTIQUEMENT MODIFIÉES

S508Q02

Dans l'étude scientifique mentionnée par l'article, quels sont les facteurs qu'on a volontairement fait varier ? Entourez « Oui » ou « Non » pour chacun des facteurs suivants.

Est-ce que, dans l'étude, on a volontairement fait varier ce facteur ?	Oui ou Non ?
Le nombre d'insectes dans l'environnement	Oui / Non
Les types d'herbicide utilisés	Oui / Non

CULTURES GÉNÉTIQUEMENT MODIFIÉES : CONSIGNES DE CORRECTION Q 2

Crédit complet

Code 1 : Les deux réponses sont correctes. Dans l'ordre : Non, Oui.

Pas de crédit

Code 0 : Autres réponses.

Code 9 : Omission.

Question 3 : CULTURES GÉNÉTIQUEMENT MODIFIÉES**S508Q03**

On a semé du maïs dans 200 champs à travers le pays. Pourquoi les scientifiques ont-ils utilisé plus d'un site ?

- A Afin que de nombreux agriculteurs puissent essayer le nouveau maïs OGM.
- B Pour voir quelle quantité de maïs OGM ils pourraient cultiver.
- C Pour recouvrir le plus de terrain possible avec des cultures OGM.
- D Pour inclure diverses conditions de culture du maïs.

CULTURES GÉNÉTIQUEMENT MODIFIÉES : CONSIGNES DE CORRECTION Q 3***Crédit complet***

Code 1 : D. Pour inclure diverses conditions de culture du maïs.

Pas de crédit

Code 0 : Autres réponses.

Code 9 : Omission.

Question 4 : CULTURES GÉNÉTIQUEMENT MODIFIÉES**S508Q04 – 0 1 9**

On a semé du maïs OGM traité avec le nouvel herbicide puissant sur la moitié de chaque champ et on a semé du maïs traditionnel traité avec de l'herbicide traditionnel sur l'autre moitié.

En quoi le fait d'utiliser chaque champ de cette façon rend-t-il l'étude objective ?

.....

.....

.....

CULTURES GÉNÉTIQUEMENT MODIFIÉES : CONSIGNES DE CORRECTION Q 4***Crédit complet***

Code 1 : La réponse doit montrer que l'élève est conscient de la nécessité de contrôler d'autres facteurs (le climat, l'hydrographie, le sol, etc.), afin d'assurer une égale représentation de la diversité des conditions de culture pour le maïs OGM et non OGM.

- Les cultures poussent sur un même sol et dans des conditions climatiques identiques.
- De manière à ce que les deux cultures aient des conditions de culture égales.
- Pour qu'ils aient un groupe de contrôle.
- Parce qu'on leur a donné la même surface de terrain et position.
- De cette manière ils peuvent dire que l'endroit n'affecte pas l'étude. [Cette réponse se réfère au fait que l'étude incluait « 200 champs à travers le pays ».]

Pas de crédit

Code 0 : Autres réponses.

- Pour les rendre comparables. *[Pas assez spécifique.]*
- En montrant qu'ils poussent dans deux (ou plusieurs) types de conditions. *[La réponse fait référence à l'utilisation de plusieurs champs mais n'identifie en aucune manière le fait que cela permet de comparer les deux traitements dans des conditions différentes.]*
- Les deux moitiés ont été cultivées de manière différente, les différences sont donc bien visibles.

Code 9 : Omission.

Question 10N : CULTURES GÉNÉTIQUEMENT MODIFIÉES

S508Q10N

Êtes-vous intéressé(e) par les informations complémentaires suivantes ?

Ne cochez qu'une case par ligne.

	<i>Cela m'intéresse beaucoup</i>	<i>Cela m'intéresse moyennement</i>	<i>Cela m'intéresse peu</i>	<i>Cela ne m'intéresse pas</i>
a) Apprendre de quelle manière on modifie génétiquement les plantes.	☐ ₁	☐ ₂	☐ ₃	☐ ₄
b) Apprendre pourquoi certaines plantes résistent aux herbicides.	☐ ₁	☐ ₂	☐ ₃	☐ ₄
c) Mieux comprendre la différence entre croiser des plantes et les modifier génétiquement.	☐ ₁	☐ ₂	☐ ₃	☐ ₄

البيت الزجاجي

اقرأ النص التالي، ثم أجب عن الأسئلة التي تليه.

تأثير البيت الزجاجي: حقيقة أم خيال؟

الكائنات الحية بحاجة إلى الطاقة لتبقى على قيد الحياة. وهذه الطاقة التي تحفظ الحياة على الأرض مصدرها الشمس، فهي تشع الطاقة إلى الفضاء لأنها حارة جداً. ونسبة صغيرة جداً من هذه الطاقة تصل إلى الأرض.

يعمل الغلاف الجوي للأرض كغطاء واقٍ يسلطح كوكبنا؛ حيث يمنع تقلبات درجة الحرارة التي يمكن أن تحدث في عالم خال من الهواء.

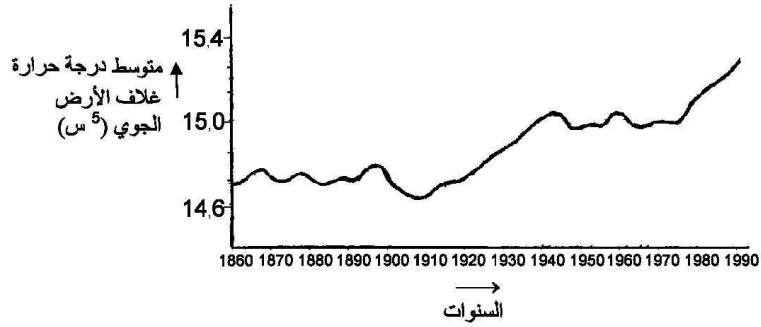
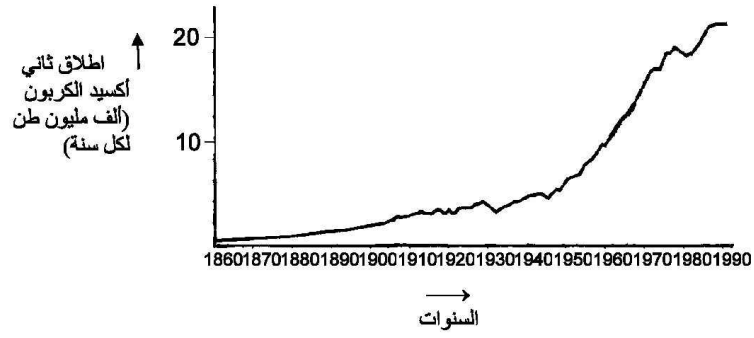
تعتبر معظم طاقة الإشعاع القادمة من الشمس الغلاف الجوي للأرض. تمتص الأرض جزءاً من هذه الطاقة و ينعكس جزء آخر عن سطح الأرض، حيث يقوم الغلاف الجوي بامتصاص جزء من هذه الطاقة المنعكسة.

ونتيجة لذلك فإن معدل درجة الحرارة فوق سطح الأرض يكون أعلى مما سيكون عليه في حالة عدم وجود غلاف جوي. إن غلاف الأرض الجوي له نفس تأثير البيت الزجاجي، ومن هنا أخذ مصطلح تأثير ظاهرة البيت الزجاجي.

وقد قيل إن تأثير ظاهرة البيت الزجاجي أصبح أكثر وضوحاً في القرن العشرين.

لقد أصبح واضحاً أن متوسط درجة حرارة الغلاف الجوي للأرض قد زاد. ويُذكر في الصحف والمجلات أن السبب الرئيسي لارتفاع درجة حرارة الغلاف الجوي في القرن العشرين هو الزيادة في إطلاق ثاني أكسيد الكربون.

اهتم أحد الطلاب واسمه خالد بدراسة العلاقة المتوقعة بين متوسط درجة حرارة الغلاف الجوي وإطلاق ثاني أكسيد الكربون على الأرض. وقد عثر على الرسمين الآتيين في إحدى المكتبات.



استنتج خالد من هذين الرسمين أن الزيادة في متوسط درجة حرارة الغلاف الجوي للأرض عائدة إلى الزيادة في إطلاق ثاني أكسيد الكربون.

S114Q03-01 02 11 12 99

سؤال30: البيت الزجاجي

من الرسمين السابقين ما الذي يدعم النتيجة التي توصل إليها خالد؟

.....
.....

S114Q04-0 1 2 9

سؤال31: البيت الزجاجي

لا توافق طالبة اسمها لين على استنتاج خالد. حيث قارنت الرسمين وقالت بأن بعض أجزاء المنحنيين لا تدعم استنتاجه.

أعط مثلاً على أجزاء من المنحنيين لا تدعم استنتاج خالد. فسر إجابتك

.....
.....
.....

S114Q05- 01 02 03 11 12 99

سؤال32: البيت الزجاجي

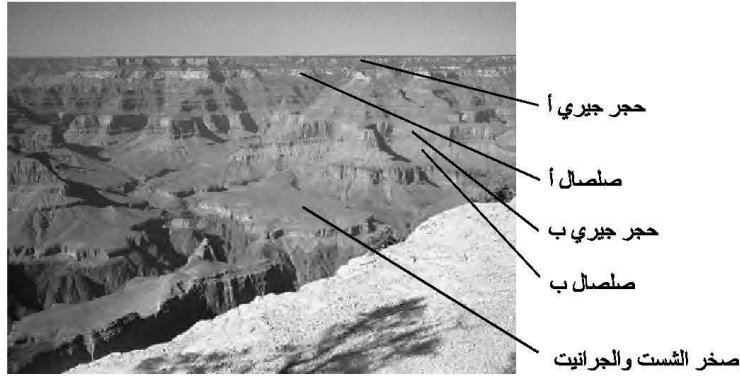
أصر خالد على استنتاجه وهو أن زيادة متوسط درجة حرارة الغلاف الجوي للأرض سببها الزيادة في إطلاق ثاني أكسيد الكربون، لكن لين كانت تعتقد أن استنتاجه غير مكتمل، وتقول: " قيل قبول هذا الاستنتاج يجب التأكد من ثبات العوامل الأخرى التي يمكن أن تؤثر على فعالية البيت الزجاجي ".
اذكر واحداً من العوامل التي قصدها لين.

.....

الوادي الكبير

يقع الوادي الكبير في صحراء في الولايات المتحدة الأمريكية. إنه واد كبير وعميق جداً يحتوي على العديد من الطبقات الصخرية. في وقت ما في الماضي أدت حركات في القشرة الأرضية إلى رفع هذه الطبقات للأعلى. ويصل عمق الوادي الآن في بعض المناطق إلى 1.6 كم. يجري نهر كولورادو عبر قاع هذا الوادي.

انظر إلى صورة الوادي الكبير أدناه، التي التقطت من حافة الوادي الجنوبية. يمكن مشاهدة العديد من الطبقات الصخرية المختلفة في جدران الوادي.



S426Q07

سؤال 21: الوادي الكبير

يزور نحو خمسة ملايين شخص المتنزه الوطني في الوادي الكبير كل سنة. هناك قلق حول التلف والخراب الذي يلحق بالمتنزه من قبل هذا العدد الكبير من الزوار. هل يمكن الإجابة عن الأسئلة الآتية بالاستقصاء العلمي؟ ضع دائرة حول "نعم" أو "لا" لكل سؤال.

هل يمكن الإجابة عن هذا السؤال بالاستقصاء العلمي؟	نعم / لا؟
ما مقدار التآكل الناتج من استعمال طرق المشاة؟	نعم / لا
هل منطقة المتنزه بنفس الجمال الذي كانت عليه قبل 100 سنة؟	نعم / لا

S426Q03

سؤال 22: الوادي الكبير

- يتراوح مدى درجات الحرارة في الوادي الكبير بين أقل من 50°C إلى ما فوق 40°C س. وبالرغم من أنها منطقة صحراوية، إلا أن الشقوق في الصخور تحتوي أحياناً على ماء. كيف تساعد تغيرات درجات الحرارة، والماء في شقوق الصخور على تسريع تكسير الصخور؟
- أ. يذيب الماء المتجمد الصخور الدافئة.
 - ب. يدعم الماء تماسك الصخور معاً.
 - ج. يُثَغِّم الجليد سطح الصخور.
 - د. يتمدد الماء المتجمد في شقوق الصخور.

S426Q05

سؤال 23: الوادي الكبير

هناك العديد من الأحافير لحيوانات بحرية مثل الرخويات، والسمك والمرجان في طبقة الحجر الجيري أ من الوادي الكبير. ما الذي حدث قبل ملايين السنين ويفسر سبب وجود هذه الأحافير هناك؟

- أ. في الأزمنة القديمة، أحضر الناس مأكولات بحرية إلى تلك المنطقة من المحيط.
- ب. كانت المحيطات في السابق أكثر هيجاناً، وأقيت الحياة البحرية على اليابسة بفعل أمواج عملاقة.
- ج. غطى محيط هذه المنطقة في ذلك الوقت ثم انسحب بعد ذلك.
- د. عاشت بعض الحيوانات البحرية يوماً ما على الأرض قبل هجرتها إلى البحر.

S426Q10S

سؤال 24: الوادي الكبير

إلى أي درجة تتفق مع العبارات الآتية؟

ضع إشارة (✓) واحدة فقط في مربع واحد فقط أمام كل عبارة.

موافق بشدة	موافق	غير موافق بشدة	غير موافق
1□	2□	3□	4□
أ. الدراسة المنظمة للأحافير مهمة.			
ب. يجب أن تستند جهود حماية المحميات الطبيعية من الضرر على الدليل العلمي.			
ج. الاستقصاء العلمي للطبقات الجيولوجية مهم.			

واقبات الشمس

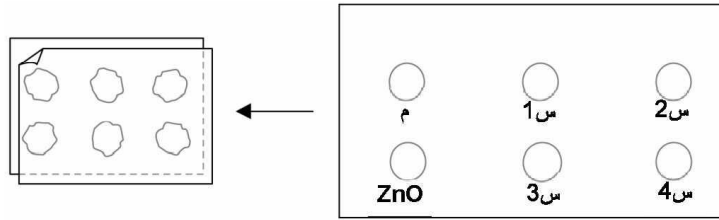
تساءلت مي وداني عن أي واقبات الشمس تقدم أفضل حماية لجلديهما. لكل منتج من منتجات واقبات الشمس معامل حماية من الشمس يرمز له بالرمز (م ح) والذي يدل على فاعلية المنتج في امتصاص الأشعة فوق البنفسجية في ضوء الشمس. المنتج الذي له (م ح) عال يحمي الجلد لمدة أطول من منتج له (م ح) منخفض.

فكرت مي في طريقة لمقارنة بعض منتجات الواقبات المختلفة. فجمعت هي وداني ما يأتي :

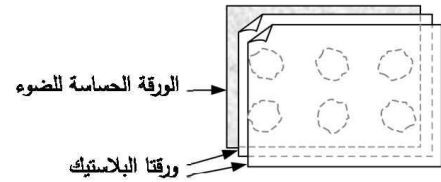
- ورقتين من البلاستيك الشفاف الذي لا يمتص ضوء الشمس ،
- ورقة واحدة حساسة لضوء الشمس ،
- زيت معدني (م)، وكريم يحتوي أكسيد الزنك (ZnO) ، و
- أربعة من منتجات واقبات الشمس المختلفة التي سمّاها س1، س2، س3، س4.

استخدمت مي وداني الزيت المعدني لأن معظم ضوء الشمس يمر من خلاله، وأكسيد الزنك لأنه يحجب تقريباً ضوء الشمس تماماً.

وضع داني قطرة من كل مادة داخل دائرة رسمت على إحدى ورقتي البلاستيك ثم وضع ورقة البلاستيك الأخرى فوقها، ثم وضع كتاباً كبيراً فوق كلتا الورقتين وضغط للأسفل.



وضعت مي ورقتي البلاستيك فوق ورقة حساسة للضوء. يتغير لون الورقة الحساسة للضوء من الرمادي الداكن إلى الأبيض (أو الرمادي الفاتح جداً) اعتماداً على مدة تعرضها لضوء الشمس. وأخيراً وضع داني الأوراق في مكان مشمس.



S447Q02

سؤال 24: واققيات الشمس

أي واحدة من العبارات الآتية تمثل الوصف العلمي لوظيفة الزيت المعدني وأكسيد الخارصين في مقارنة فاعلية واققيات الشمس؟

- أ- الزيت المعدني وأكسيد الخارصين عاملان يتم اختبارهما.
- ب- الزيت المعدني هو عامل يتم اختباره وأكسيد الخارصين مادة ضابطة.
- ج- الزيت المعدني مادة ضابطة وأكسيد الخارصين عامل يتم اختباره.
- د- الزيت المعدني وأكسيد الخارصين كلاهما مادتان ضابطتان.

S447Q03

سؤال 25: واققيات الشمس

أي واحد من الأسئلة الآتية حاول كل من مي وداني إجابته؟

- أ- كيف تقارن بين مقدار حماية كل واق مع حماية الواققيات الأخرى؟
- ب- كيف تعمل واققيات الشمس على حماية جلدك من الأشعة فوق البنفسجية؟
- ج- هل هناك من الواققيات المعطاة من هو أقل حماية من الزيت المعدني؟
- د- هل هناك من الواققيات المعطاة من هو أكثر حماية من أكسيد الخارصين؟

S447Q04

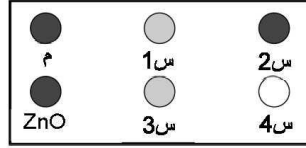
سؤال 26: واققيات الشمس

لماذا ضغطت ورقة البلاستيك الثانية للأسفل؟

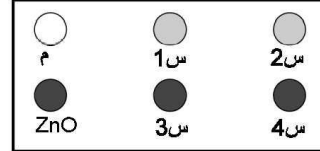
- أ. لمنع جفاف القطرات.
- ب. لتنتشر القطرات على أكبر مساحة ممكنة.
- ج. لحفظ القطرات داخل الدوائر المرسومة.
- د. لجعل القطرات متماثلة في السمك.

سؤال 27: واقيات الشمس

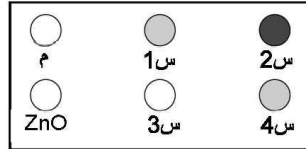
لون الورقة الحساسة للضوء رمادي داكن يتحول إلى رمادي فاتح بتعرضه لكمية قليلة من الضوء، ثم يتحول إلى اللون الأبيض عند تعرضه إلى كمية كبيرة من الضوء.
أي واحد من الأشكال الآتية، يبين نموذجاً لما يمكن أن يحدث للورقة الحساسة للضوء؟
فسّر سبب اختيارك له.



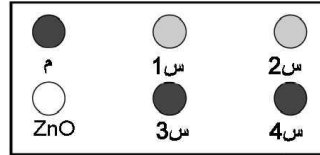
ج -



ا -



د -



ب -

الإجابة:

التفسير:

محاصيل معدلة جينياً

يجب منع الذرة المعدلة جينياً

تطالب مجموعات حماية الحياة البرية بمنع الذرة الجديدة المعدلة جينياً.

صممت هذه الذرة المعدلة جينياً بحيث لا تتأثر بمبيد الأعشاب القوي الجديد الذي يقتل نباتات الذرة التقليدية. سيقتل هذا المبيد العشبي الجديد معظم الأعشاب الضارة التي تنمو في حقول الذرة.

يقول المحافظون على البيئة بما أن هذه الأعشاب الضارة غذاء لحيوانات صغيرة خاصة الحشرات، سيكون استعمال المبيد العشبي الجديد مع الذرة المعدلة جينياً مضرراً بالبيئة. يقول مؤيدو استخدام الذرة المعدلة جينياً أن دراسة علمية أظهرت أن هذا لن يحدث.

هذه تفاصيل الدراسة العلمية التي ذكرت في المقال السابق:

- زرعت الذرة في 200 حقل في كافة أنحاء البلاد.
- قسم كل حقل إلى قسمين. زرعت الذرة المعدلة جينياً المعالجة بمبيد الأعشاب القوي الجديد في أحد النصفين، وزرعت الذرة التقليدية المعالجة بمبيد أعشاب تقليدي في النصف الآخر.
- كان عدد الحشرات التي وجدت في الذرة المعدلة جينياً المعالجة بمبيد الأعشاب الجديد مساوياً تقريباً لعدد الحشرات التي وجدت في الذرة التقليدية التي عولجت بمبيد أعشاب تقليدي.

S508 Q02

سؤال 13: محاصيل معدلة جينياً

ما العوامل التي تم تغييرها عمداً في الدراسة العلمية الواردة في المقالة؟
ضع دائرة حول "نعم" أو "لا" لكل من العوامل التالية.

هل هذا العامل تم تغييره بقصد في الدراسة؟	نعم أو لا؟
عدد الحشرات في البيئة	نعم / لا
نوع مبيد الأعشاب المستعمل	نعم / لا

سؤال 14: محاصيل معدلة جينياً

- زرعت الذرة في 200 حقل في كافة أنحاء البلاد. لماذا استعمل العلماء أكثر من موقع واحد؟
- حتى يتمكن العديد من المزارعين أن يجربوا الذرة المعدلة جينياً.
 - لمعرفة كم يمكن أن يزرعوا من الذرة المعدلة جينياً.
 - لتغطية أكبر عدد ممكن من الأراضي بالمحصول المعدل جينياً.
 - لتشمل الدراسة ظروف نمو متنوعة للذرة.

سؤال 16: محاصيل معدلة جينياً

ما درجة اهتمامك بالمعلومات الآتية؟

ضع إشارة (✓) واحدة فقط في مربع واحد فقط أمام كل معلومة.

اهتمام عالي	اهتمام متوسط	اهتمام منخفض	غير مهتم
☐ 1	☐ 2	☐ 3	☐ 4
أ- تعلم العمليات التي يتم بها تعديل النباتات جينياً			
ب- تعلم لماذا لا تتأثر بعض النباتات بمبيدات الأعشاب			
ج- فهم أفضل للاختلافات بين التهجين والتعديل الجيني للنباتات			

Appendix 3.2 – ITC Guidelines for Test Adaptation

Context

- C.1 (1) Effects of cultural differences that are not important to the main purposes of the study should be minimized to the extent possible.
- C.2 (2) The amount of overlap in the construct measured by the test in the populations of interest should be assessed.

Test development and adaptation

- D.1 (3) Test developers/publishers should ensure that the adaptation process takes full account of linguistic and cultural differences in the intended populations.
- D.2 (4) Test developers/publishers should provide evidence that the language used in the test directions, scoring rubrics, and the items themselves are appropriate for all cultural and language populations for whom the test is intended.
- D.3 (5) Test developers/publishers should provide evidence that the choice of testing techniques, item formats, test conventions, and other procedures are familiar to all intended populations.
- D.4 (6) Test developers/publishers should provide evidence that item content and stimulus materials are familiar to all intended populations.
- D.5 (7) Test developers/publishers should compile judgemental evidence, both linguistic and psychological, to improve the accuracy of the adaptation process and compile evidence on the equivalence of all language versions.
- D.6 (8) Test developers/publishers should ensure that the data collection design permits the use of appropriate statistical techniques to establish construct and item equivalence among the language versions of the test.
- D.7 (9) Test developers/publishers should apply appropriate statistical techniques to (a) establish the equivalence of the language versions of the test, and (b) identify problematic components or aspects of the test that may be inadequate in one or more of the intended populations.
- D.8 (10) Test developers/publishers should provide information on the validity of the adapted versions of the test in the intended populations.
- D.9 (11) Test developers/publishers should provide statistical evidence about the equivalence of items in all intended populations.
- D.10 (12) Non-equivalent items across the intended populations should not be used in “linking” adapted versions of the test to a common score reporting scale. However, these same items may be useful for reporting scores in each population, separately.

Administration

- A.1 (13) Those aspects of the environment that influence the administration of a test should be made as similar as possible across populations for whom the test is intended.

- A.2 (14) Test developers and administrators should try to anticipate the types of problems that can be expected, and take appropriate actions to remedy these problems through the preparation of appropriate materials and instructions.
- A.3 (15) Test administrators should be sensitive to a number of factors related to the stimulus materials, administration procedures, and response modes that can moderate the validity of the inferences drawn from the scores.
- A.4 (16) Test administration instructions should be in the source and target languages to minimize the influence of unwanted sources of variation across populations.
- A.5 (17) The test manual should specify all aspects of the test and its administration that require scrutiny in the application of the test in a new cultural context.
- A.6 (18) The administrator should be unobtrusive and the administrator-examinee interaction should be minimized. Explicit rules that are described in the test administration manual should be followed.

Documentation/Score Interpretations

- I.1 (19) When a test is adapted for use in another population, documentation of the changes should be provided, along with evidence to support the equivalence of the adapted version of the test.
- I.2 (20) Score differences among samples of populations administered the test should not be taken at face value. The researcher has the responsibility to substantiate the meaningfulness of the differences with other empirical evidence.
- I.3 (21) Comparisons across populations can only be made at the level of invariance that has been established for the scale on which scores are reported.
- I.4 (22) The test developer should provide specific information on the ways in which the socio cultural and ecological contexts of the populations might affect performance on the test, and should suggest procedures to account for these effects in the interpretation of results.

Appendix to 6.1 – Recoding PISA data in RUMM2030

M/R: multiple responses; N/A : Not available

**Full credit responses vary in position.*

Dichotomous items

PISA codes	Meaning	RUMM2030 codes	RUMM2030 score
0	No credit	0	0
1	Full credit*	1	1
8	M/R	0	0
9	Embedded missing	0	0
r	Unreached	*	*
7	N/A	*	*

Partial Credit Items (polytomous with 3 response categories)

PISA codes	Meaning	RUMM2030 codes	RUMM2030 score
0	No credit	0	0
1	Partial	1	1
2	Full credit*	2	2
9	Embedded missing	0	0
r	Unreached	*	*
7	N/A	*	*

Multiple Choice Items

Three alternatives

PISA codes	Meaning	RUMM2030codes	RUMM2030 score
0	No credit	0	0
1	No credit	1	0
2	Full credit	2	1
8	M/R	3	0
9	Embedded missing	4	0
r	Unreached	*	*
7	N/A	*	*

Four alternatives (starting with PISA code 0)

PISA codes	Meaning	RUMM2030 codes	RUMM2030 score
0	No credit	0	0
1	No credit	1	0
2	No credit	2	0
3	Full credit	3	1
8	M/R	4	0
9	Embedded missing	5	0
r	Unreached	*	*
7	N/A	*	*

Four alternatives (starting with PISA code 1)

PISA codes	Meaning	RUMM2030codes	RUMM2030 score
1	No credit	1	0
2	Full credit	2	1
3	No credit	3	0
4	No credit	4	0
8	M/R	5	0
9	Embedded missing	6	0
r	Unreached	*	*
7	N/A	*	*

Five alternatives (starting with PISA code 0)

PISA codes	Meaning	RUMM2030codes	RUMM2030 score
0	No credit	0	0
1	No credit	1	0
2	No credit	2	0
3	No credit	3	0
4	Full credit	4	1
8	M/R	5	0
9	Embedded missing	6	0
r	Unreached	*	*
7	N/A	*	*

Five alternatives (starting with PISA code 1)

PISA codes	Meaning	RUMM2030codes	RUMM2030 score
1	No credit	1	0
2	No credit	2	0
3	No credit	3	0
4	Full credit	4	1
5	No credit	5	0
8	M/R	6	0
9	Embedded missing	7	0
r	Unreached	*	*
7	N/A	*	*

Six alternatives (starting with PISA code 0)

PISA codes	Meaning	RUMM2030codes	RUMM2030 score
0	No credit	0	0
1	No credit	1	0
2	No credit	2	0
3	No credit	3	0
4	No credit	4	0
5	Full credit	5	1
8	M/R	6	0
9	Embedded missing	7	0
r	Unreached	*	*
7	N/A	*	*

Item S485Q02 SCIE-P2006 Acid Rain

PISA codes	Meaning	RUMM2030 codes	RUMM2030 score
0	No credit	0	0
1	Full credit	1	1
2	Full credit	2	1
9	Embedded missing	3	0
r	Unreached	*	*
7	N/A	*	*

Appendix to 7.1 – DIF items (Country curves)

