Title
Differential Item Functioning: The Consequence of Language, Curriculum, or Culture?

Author
Huang, Xiaoting

Publication Date
2010

Peer reviewed|Thesis/dissertation

Differential Item Functioning:
The Consequence of Language, Curriculum, or Culture?

By

Xiaoting Huang

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy
in
Education
in the
Graduate School of Education
of the
University of California, Berkeley

Committee in charge:
Professor Mark Wilson
Professor Sophia Rabe-Hesketh
Professor Carla Hudson Kam

Fall 2010

Abstract

Differential Item Functioning: the Consequence of Language, Curriculum or Culture?

By

Xiaoting Huang

Doctor of Philosophy in Education

University of California, Berkeley

Professor Mark Wilson, Chair

In recent decades, the use of large-scale standardized international assessments has increased drastically as a way to evaluate and compare the quality of education across countries. In order to make valid international comparisons, the primary requirement is to ensure the measurement equivalence between the different language versions of these assessments due to their multilingual and cross-cultural nature. In this study, we investigated the measurement equivalence of one of the most popular international assessments, PISA (Programme for International Student Assessment), between U.S. and Canadian, Hong Kong and mainland Chinese, and U.S. and mainland Chinese students. Both unidimensional and multidimensional random coefficient multinomial logit model (RCML) were applied to detect differential item functioning (DIF). Furthermore, we exerted great efforts to identify possible explanations of DIF via detailed content analyses. The results showed that the number of DIF items is the smallest between Canadian and U.S. students and the largest between U.S. and Chinese students. We also noticed that for all three comparisons the number of DIF items reduced significantly when we analyzed the data using the multidimensional approach. Our content analysis revealed that language difference only accounted for a small proportion of DIF between U.S. and Chinese students, whereas differential curriculum coverage was found to be the most serious cause of DIF in both the Hong Kong-Mainland and the U.S.-Chinese comparisons. In addition, we found that differential content familiarity is also a potential cause of DIF. Further investigations of more potential sources of item bias require the collection of additional data.

Dedication

To my wonderful family, for their endless encouragement, support and love, this dissertation is dedicated.

Acknowledgements

First of all, I want to thank my committee chair and academic advisor, Professor Mark Wilson, for his wonderful mentorship and guidance throughout my graduate years at Berkeley. Mark is very passionate about research. His enthusiasm greatly encouraged me to pursue my own interests and professional goals. And Mark is always supportive, giving me freedom to explore my own interests while providing guidance throughout all challenges. Without him, the completion of this work would not be possible.

I would also like to express my gratitude to my second advisor, Prof Sophia Rabe-Hesketh, for all her encouragement and support. Sophia is also an amazing role model for me. She is always energetic and works very diligently. She set an excellent example of the kind of educator and researcher I hope to become in the future.

I am definitely thankful to my third committee member, Prof Carla Hudson Kam, for her support on the completion of this dissertation.

I would also like to thank Prof Wen-Chung Wang from the Hong Kong Institute of Education for his insightful suggestions and comments on this study as well as all his kind help during my Ph.D. years.

My special thanks go to Prof. Ji Shen from the University of Georgia, Ms. Lei Wang from NEEA (China National Education Examinations Authority), and Ms. Xuelan Qiu from the Hong Kong Institute of Education, for their enormous help with my investigations on the curriculum differences among U.S., Hong Kong and mainland China.

Finally, my deepest gratitude goes to my family, my parents and my husband, for their understanding, support and unconditional love. To them I dedicate this dissertation.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Background

The popularity of international large-scale standardized testing

In recent decades, with the rapid growth of globalization, the demands for cross-national cooperation have become higher than ever. Tom Friedman's idea that "the world is flat" has won loud cheers internationally (Friedman, 2005). Waving this same banner, the use of large-scale standardized international assessments has increased drastically as a way to evaluate and compare the quality of the future labor force across countries. In fact, many countries spend a large amount of money on international student assessments, believing that those assessments will identify policy solutions to the shortcomings in their education systems.

Currently, there are three major international student assessments, known widely by their acronyms.

1. **PIRLS** (Progress in International Reading Literacy Study), developed by IEA (International association for Evaluation of Educational achievement), is an assessment designed to measure trends in 4th grade children's reading achievement and policy and practices related to literacy. It was first conducted in 2001 and was carried out every 5 years since then. In 2006, 40 countries and jurisdictions participated in the assessment. (http://www.iea.nl/pirls2011.html)

2. **TIMSS** (Trends in International Mathematics and Science Study), another major project of IEA, is an assessment designed to measure students' science and math achievements in 4th and 8th grades, and on related contextual aspects such as mathematics and science curricula and classroom practices across countries. It was conducted on a 4-year cycle starting from 1995. In 2007, the fourth cycle of TIMMS, 36 countries at grade four and 48 countries at grade eight participated. 33 countries participated in both assessments. (http://www.iea.nl/timss2011.html)

3. **PISA** (Programme for International Student Assessment), developed by OECD (Organization for Economic Co-operation and Development), is an assessment of reading, mathematical, and scientific "literacy" among 15-year-olds. It was administered every three years since 2000. In 2006, 57 countries and jurisdictions took part in the assessment. (http://nces.ed.gov/Surveys/PISA/)

These assessments have collected huge datasets from thousands of participants all over the world, providing valuable information about students' cognitive development or academic achievement, as well as key demographic, economic, social and educational information that may influence students' performance. These databases are widely used

to make interesting international comparisons, for policy analysis and all sorts of research purposes.

Introduction of PISA

Among all the well-known international assessments, PISA is unique in several ways. First, unlike TIMSS or PIRLS, which are more grade- and curriculum-centered, PISA proclaims to assess the "literacy" of 15-year-olds approaching the end of compulsory schooling around the world, and is not tied to any specific curricula. Rather than focusing on the extent to which these students have mastered a specific school curriculum, it looks at their ability to use their knowledge and skills to meet real-life challenges. It also has an emphasis on globalization and claims to be assessing the skills that young adults will need in the emerging global economy in the 21st century. (OECD, 2006)

Secondly, PISA collects student achievement information in more subject areas than any other existing international assessments. It is a triennial survey of the knowledge and skills of *Reading*, *Mathematical* and S*cientific Literacy*. (OECD, 2006) Four PISA surveys have taken place so far, in 2000, 2003, 2006 and 2009. It collects information in all three domains through three-year cycles with emphasis in one major domain each cycle.

In addition, PISA has the widest participation. The high visibility of the OECD has propelled PISA forward to become the most visible of the international assessments. More than 400,000 students from 57 countries and areas making up close to 90% of the world economy took part in PISA 2006 (OECD, 2007). (Although not officially listed, Mainland China also participated in it for the first time in 2006.) In 2009, 66 countries and economies participated in the assessment. As a result, PISA can be viewed as the product of collaboration between all participating countries and economies through OECD. It represents a commitment by the governments of OECD member countries to monitor the outcomes of education systems in terms of student achievement, within a common international framework. It draws on leading international expertise to develop valid measures for making comparisons across countries and cultures (OECD, 2007).

The use of international assessments for cross-country comparisons and the associated test equivalence problem

Given the worldwide reach of international assessments like PISA, it is hard not to view the collected data as a rich source for cross-country comparisons. In fact, it is one of the most important purposes of all international assessments. Otherwise, there would be no need for the involvement of countries other than one's own.

"International benchmarking" (Helgason, 1997) is the term adopted to describe the numerous efforts in using those assessment results for policy purposes. It has at least

two components: the first is to monitor and evaluate the current educational quality by comparing the relative performance of students in one country versus their peers in other countries. The second function is to identify effective practices, usually from high-performing countries, to provide policy advice on how to improve student performance in the future. By linking the collected contextual information with student achievement data, researchers and policy makers try to address various questions like what student-level factors (such as gender and social background) and school-level factors may improve achievement; what the cross-countries differences are in the relationship between those factors and achievement; and how countries differ in education systems and national contexts that are related to differences in student achievement, just to name a few (Afonso & St. Aubyn, 2006; Enkins, 2006; Fuchs & Wößmann, 2006).

While we enthusiastically embrace the enormous pool of information provided by these international assessments, the limitations of the obtained comparative results are often overlooked. One important feature of these assessments is that they are multilingual and cross-cultural in nature. It is of primary concern to ensure the functional equivalence of these assessments in order to make valid international comparisons. One requirement for functional equivalence is that all dimensions of the assessment (e.g. student achievement, student self-reported interest, schools' educational resources, etc) should hold the same meaning across languages and cultures.

The equivalence of test items can be compromised when translation flaws or differences in students' familiarity with the question content or context cause item-by-country interactions, that is, an item appears to be easier or harder for students at a same level of proficiency in different countries. When an international assessment contains a considerable amount of such items, the interpretation of the results becomes problematic, because the instability of item difficulties across countries prevents accurate descriptions of the skills associated with the proficiency estimates.

The construction of international assessments usually involves rules and procedures designed to guarantee the literal equivalence of the different language versions used within and between participating countries, and to take into account the diverse cultural contexts of all participating societies. However, many studies have shown that the comparability of test scores between different language versions of those tests is often at risk despite all the precautions (Allalouf, 2000; Grisaya & Monseur, 2007; Price & Oshima, 1998; Sireci & Swaminathan, 1996). Hence, this issue of "measurement equivalence" and how to address it becomes a fundamental challenge for international assessments.

1.2 DIF Analysis in International Large-Scale Assessments

What is DIF?

Differential item functioning (*DIF*) (Holland & Wainer, 1993) analysis is an especially suitable and sharp tool for assessing the functional equivalence in different

versions of international assessments. It is an item response theory (IRT) based approach to evaluate the measurement equivalence of assessments. In fact, it has become a routine practice of item analysis in many commercial tests, especially in high-stakes tests.

An item is said to exhibit DIF when it functions differently for different groups of test-takers controlling for their different abilities. Note that it is not the group difference in the measured performance by the test or item. DIF occurs when test-takers, who have identical levels of a certain latent trait a test was designed to measure but belong to different subpopulations, have significantly different probabilities of answering a particular item correctly (or endorsing an item). For example, when an item that is intended to measure mathematical proficiency includes a slang term that is only known to black students, students from other ethnicity groups would have a lower probability of answering that item correctly, even though they possess equal proficiency in mathematics. Then this item is viewed as biased against other subgroups of examinees.

In a typical DIF study, the item responses of two groups of test-takers are examined: a reference group, which is often the majority group, and a focus group, which is often a minority group. The grouping variable is usually bi-categorical, but it can be multi-categorical as well (Wang, 2008). To investigate the measurement equivalence in different language versions of large-scale international assessments, such as PISA, studies often examine the test invariance across countries using country membership as the grouping variable. When a test contains many DIF items, the cross-country comparability is at risk. The interpretation of a scale can be severely biased due to unstable item characteristics from one country to another.

DIF studies on large-scale international assessments

Quite a few studies have found the presence of DIF in cross-language cross-country assessments (Allalouf, 2000; Budgell, Raju, & Quartetti, 1995; Ercikan & Koh, 2005; Grisaya & Monseur, 2007; Price & Oshima, 1998; Sireci & Swaminathan, 1996; A. D. Wu & Ercikan, 2006). For example, Wu and Ercikan (2006) identified DIF between Taiwan and the United States in the TIMSS 1999 test. Ercikan and Koh (2005) examined the equivalence of English and French versions of TIMSS 1995 test and found large numbers of DIF items. Budgell, Raju, and Quartetti (1995) reviewed studies of DIF in translated assessment instruments and found that the number of items with DIF ranged from 1.5% to 64%, with many studies finding DIF in over 30% of the items. These studies show that DIF prevails in international assessments, posing serious threat to the "fairness" of those instruments.

Although the statistical procedures of detecting DIF have been discussed frequently (Clauser & Mazor, 1998; Ronald K. Hambleton & Rogers, 1989; Shealy & Stout, 1993; Wang, 2008), efforts to identify and understand the causes of DIF are scarce, especially for cross-country DIF in international assessments. Among the few studies that are devoted to exploring the causes of DIF, the majority focuses on the fidelity of test translation (Allalouf, Hambleton, & Sireci, 1999; Ercikan, 1998; R. K. Hambleton &

Kanjee, 1995; R. K. Hambleton & Patsula, 1999). Others suggest that there are many possible reasons why item bias may occur, for example, inadequate test administration, statistical artifacts (floor or ceiling effects), differential familiarity with the stimulus materials or response formats, tapping different traits in various groups, etc (Vijver & Poortinga, 1985). Only a few studies go beyond these factors to investigate cultural sources of DIF (A. D. Wu & Ercikan, 2006).

The findings of these studies are valuable first steps to locate the sources of DIF. But they are far from comprehensive or conclusive. Identifying the causes of DIF in international assessments is especially challenging because such causes are often nebulous and intertwined. Detailed item analysis on a larger item pool is therefore called for to provide deeper insights into the actual causes of DIF as well as possible ways to revise DIF items.

1.3 The DIF Problem in the PISA 2006 Assessment

Efforts to ensure PISA 2006 test equivalence

Test Development As the aim of PISA is to develop reliable "literacy" scales for all participating countries and provide fully comparable information, OECD's test development team undertook tremendous efforts to make PISA as "fair" as possible across nations. Extra precautions were taken throughout the entire process of test development (OECD, 2009a). Specifically, for PISA 2006, test development centers were established in five culturally diverse and well-known institutions, namely ACER (Australia), CITO (the Netherlands), ILS (University of Oslo, Norway), IPN (University of Kiel, Germany) and NIER (Japan). By establishing these centers, PISA hopes to achieve conceptually rigorous material that has the highest possible levels of cross-cultural and cross-national diversity.

These test development centers were in charge of the initial development of items. The process of developing cognitive items started with a calling for submissions from participating countries. A total of 155 item units were processed from 21 countries. Each submitted unit was reviewed by one of the test development centers to determine its general suitability for PISA 2006. The process was then continued with local item paneling and local pilot tests. Afterwards, the items were revised, followed by international item paneling and international pilot tests.

Test Translation Translation from the two source languages, English and French, took place only after the items had reached a well-formed stage. PISA was the first major international assessment that uses two different source languages. This parallel development of the two source versions assisted in making the items as culturally neutral as possible. Both English and French source versions of all test instruments were then distributed to participating countries as a basis for local adaptation and translation into national versions. This procedure is called double translation. Unlike the traditional test

translation procedure, PISA does not involve a back translation which translates the test from the national version back to its source language and compares the back translated version with the original one.

Resorting to two different languages may, to a certain extent, reduce problems linked to the impact of cultural characteristics of a single source language. However, even though English and French do represent relatively different sets of cultural traditions, both languages share an Indo-European origin, which may have little effect for countries where the languages do not belong to the Indo-European family.

Item-by-Country Interaction Analysis After the items were field tested, PISA checked the item-by-country interaction from a quantitative approach. Specifically, the data were scaled with the Rasch model for each country and for many languages. The relative difficulty of an item *i* for a language *j* within a country *k*, denoted as d_{ijk}, was compared with its international relative difficulty $d_{i..}$, computed on a random sample from 51 participating OECD countries where each national sample was the same size. The item-by-country interaction is defined as the difference between any d_{ijk} and its corresponding international item difficulty $d_{i..}$.

As both the national and international item calibrations were centered at zero, the mean of the item difficulties, d_{ijk} , for any language *j* within a country *k* is equal to zero: $\sum_{i=1}^{I}\delta_{ijk}=0$. Therefore, the sum (and consequently the arithmetic mean) of the item-by country interaction for a particular language version within a country is always equal to zero: $\sum_{i=1}^{I}\left(\delta_{ijk}-\delta_{i..}\right)=\sum_{i=1}^{I}\delta_{ijk}-\sum_{i=1}^{I}\delta_{i..}=0$, where *I* is the number of items in the test. The mean absolute deviation of the interaction for that language version (language *j*) is: $MAD_{jk}=\frac{1}{I}\sum_{i=1}^{I}\left|\delta_{ijk}-\delta_{i..}\right|$; and the root mean squared error is: $RMSE_{jk}=\sqrt{\frac{1}{I}\sum_{i=1}^{I}\left(\delta_{ijk}-\delta_{i..}\right)^2}$. A chi-square statistic is consequently: $\chi^2=\sum_{i=1}^{I}\frac{\left(\delta_{ijk}-\delta_{i..}\right)^2}{\mathrm{var}\left(\delta_{ijk}\right)}$, with one degree of freedom.

This statistic gives an overall evaluation of the item-by-country interaction for each language version *j* in country *k*. Six versions were found to have the highest mean deviation: the Azeri version from Azerbaijan; the Uzbek version from Kyrgyzstan; the Kyrgyz version from Kyrgyzstan; the Russian version from Azerbaijan; the Hungarian version from Romania; and the Chinese version from Chinese Taipei (OECD, 2009a).

At item level, the mean absolute deviation of interaction for a particular item unit (items sharing the same stimulus material were defined as one item unit) across different language versions was also calculated, $MAD_{ik}=\frac{1}{J}\sum_{j=1}^{J}\left|\delta_{ijk}-\delta_{i..}\right|$, where *J* is the number of

language versions. On average across item units, the average interaction is 0.34, ranging from 0.25 to 0.44. Item units with large interactions were either removed or revised (OECD, 2009a).

PISA allows an item may be deleted (coded as not administered) from national calibration if it has poor psychometric properties in that country, and be deleted from international calibration if it has poor psychometric characteristics in more than ten countries. In fact, a few science items were removed from the science item parameter database for national and international parameter estimates (OECD, 2009a).

Why cross-country DIF analysis is still needed for PISA 2006

The measurement equivalence of PISA can never be over emphasized since policy makers across the world are using PISA findings to gauge the knowledge and skills of students in their own country in comparison with those of other participating countries; establish benchmarks for educational improvement. Since "Science literacy" is the major domain for the first time in 2006 and the results provide the baseline for future measures of change in this subject, it is of critical importance to investigate thoroughly the measurement equivalence of this scale.

Although the development of PISA 2006 underwent arduous procedures to prevent measurement inequivalence in its final version, the comparability between specific countries remains a question. The mean absolute deviation indices for item units across all language versions only provide some hints on potential differential item functioning because it is the mean difference of difficulty estimates for each version and the pooled international average. However, there is no guarantee that the item holds relatively similar difficulties and targets similar knowledge or skills for any two specific countries. On the other hand, the mean absolute deviation statistic for a particular language version only shows whether the difficulty estimates of all items in that version differ in general from those for the international sample. In other words, we cannot locate problematic items from this statistic, let alone find the reasons of the item-by-country interactions.

Previous studies have found the existence of cross-national DIF in PISA 2006 reading and 2003 math scales (Grisaya & Monseur, 2007; Xie & Wilson, 2008). Little can be found evaluating the cross-country validity of its 2006 science scale. However, we have reasons to believe that this scale may also have flaws. First, the PISA item-by-country interaction inspection results indicate that there are countries that differ significantly from international pooled samples (OECD, 2009a). It is highly likely that comparisons between those countries with others are problematic.

Secondly, the complex nature of the concept of "scientific literacy" may lead to DIF. PISA defines the term "scientific literacy" as an individual's scientific knowledge and use of that knowledge to identify questions, to acquire new knowledge, to explain scientific phenomena, and to 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, and willingness to engage in science-related issues, and with the ideas of science, as a reflective citizen (OECD, 2009b). This term embodies the desired goals of science education reform and in turn, reflects the desired components of science education. PISA views the ability to solve problems, to communicate, and to use information technology as high priority goals for the future development of countries around the world (OECD, 1999). Unfortunately, the vision of science education and the desired goals are often diffuse and dependent upon the context and purpose within a society (Kemp, 2000). Even with similar definitions of scientific literacy, the expectations for success are different, ranging from high standards to minimum competencies. Hence, the interpretation of what was assessed by the Science scale may vary from country to country.

Therefore, DIF analysis is initiated to investigate the comparability of the PISA 2006 Science scale between specific countries.

1.4 Research Questions

As mentioned earlier in this chapter, the PISA 2006 science scale aims to provide valid comparisons across countries and cultures. To achieve that goal, it is critical to ensure the validity equivalence among students from different participating countries and areas. Furthermore, if DIF problems are detected in the PISA 2006 data, the relatively large number of items provides a great opportunity for us to seek underlying causes via detailed content analyses. The substantive analysis results will be invaluable to many stakeholders in that it may reveal information such as what different factors were actually assessed by the test for different subpopulations, and how different educational policies may be reflected in the assessment. In addition, it can also inform test developers on how to improve their future practices in terms of ensuring the "fairness" across countries by finding out what imperfections in test translation or item content adaption may hurt the functional equivalence of items.

The current study proposes to examine the cross-country validity of PISA 2006 Science scale among four groups of students. Specifically, I performed three DIF analyses: between U.S. and Canadian students, between mainland Chinese and Hong Kong students, and between U.S. and mainland Chinese students. These four groups of students offer interesting comparisons between students with the same official language, similar culture and similar curricula; similar languages, culture but different curricula; and different languages, culture and different curricula, respectively. Content analyses will be conducted following the statistical DIF detection step.

By performing these analyses, this study attempts to answer two main research questions:

(1) Is there a serious threat to the functional equivalence of the PISA 2006 Science scale for any of the three comparisons?

(2) If a substantial amount of items are detected to display DIF, what might be the underlying causes? Specifically, is DIF the consequence of test translation/adaptation, different curriculum coverage, or cultural difference?

Reference

Afonso, A., & St. Aubyn, M. (2006). Cross-country efficiency of secondary education provision: A semi-parametric analysis with non-discretionary inputs. *Economic Modelling, 23*(3), 476-491.

Allalouf, A. (2000). *Retaining Tranlated Verbal Reasoning Items by Revising DIF Items.* Paper presented at the AERA, New Orleans, LA.

Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the causes of DIF in translated ver- bal items. *Journal of Educational Measurement, 36*(3), 185–198.

Buckley, J. (2009). *Cross-National Response Styles in International Educational Assessments: Evidence from PISA 2006.* Paper presented at the NCES PISA Research Conference.

Budgell, G. R., Raju, N. S., & Quartetti, D. A. (1995). Analysis of differential item functioning in translated assessment instruments. . *Applied Psychological Measurement,, 19,* 309-321.

Cheung, S. K. (1996). Reliability and factor structure of the Chinese version of the Depression Self-Rating Scale. *Educational and Psychological Measurement, 56,* 142-154.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differential item functioning test items. *Educational Measurement: Issues and Practice, 17*(1), 31–44.

Enkins, S. P., Micklewright, John and Schnepf, Sylke V. (2006). *Social Segregation in Secondary Schools: How does England Compare with other Countries?* Paper presented at the IZA Discussion.

Ercikan (1998). Translation effects in international assessment. *International Journal of Educational Research, 29,* 543–553.

Ercikan, & Koh (2005). Examining the Construct Comparability of the English and French Versions of TIMSS. *International Journal of Testing, 5*(1), 23 - 35.

Friedman, T. (2005). *The World is Flat: A Brief History of the Twenty-First Century*. New York: Farrar, Straus & Giroux.

Fuchs, T., & Wößmann, L. (2006). What accounts for international differences in student performance? A re-examination using PISA data. *Empirical Economics, 32*(2-3), 433-464.

Grisaya, A., & Monseur, C. (2007). MEASURING THE EQUIVALENCE OF ITEM DIFFICULTY IN THE VARIOUS VERSIONS OF AN INTERNATIONAL TEST. *Studies in Educational Evaluation, 33*(1), 69-86

Hambleton, R. K., & Kanjee, A. (1995). Increasing the validity of cross-cultural assessments: Use of improved methods for test adaptations. *European Journal of Psychological Assessment, 11*, 147–157.

Hambleton, R. K., & Patsula, L. (1999). Increasing the validity of adapted tests: Myths to be avoided and guidelines for improving test adaptation practices. *Journal of Applied Testing Technology, 1,* 1-12.

Hambleton, R. K., & Rogers, H. J. (1989). Detecting Potentially Biased Test Items: Comparison of IRT Area and Mantel-Haenszel Methods. *Applied Measurement in Education, 2*(4), 313-334.

Helgason, S. (1997). *International Benchmarking: Experiences from OECD countries*, Copenhagen.

Holland, P. W., & Wainer, H. (1993). *Differential Item Functioning.* Hillsdqle, NJ: rlbaum.

Kemp, A. C. (2000). *Science educator's views on the goal of scientific literacy for all: An interpretative review of the literature*. Paper presented at the the Annual Meeting of the Nation Association for Research in Science Teaching.

OECD (1999). *Measuring student knowledge and skills: A new framework for assessment*.

OECD (2006). *Assessing Scientific, Reading and Mathematical Literacy: a Framework for PISA 2006*.

OECD (2007). *PISA 2006: Science Competencies for Tomorrow's World* (Vol. 1).

OECD (2009a). *PISA 2006 Techinical Report*.

OECD (2009b). PISA Data Analysis Manual: SAS Second Edition.

Price, L. R., & Oshima, T. C. (1998). *Differential Item Functioning and Language Translation: A Cross-National Study with a Test Developed for Certification.* Paper presented at the AERA, San Diego, CA.

Shealy, R., & Stout, W. (1993). An item response theory model for test bias and differential item functioning. In P. W. Holland & W. Howard (Eds.), *Differential item functioning*. Hillsdale: NJ: Lawrence Erlbaum.

Sireci, S. G., & Swaminathan, H. (1996). *Evaluating Translation Equivalence: So What's the Big DIF?* Paper presented at the AERA, Ellenville, NY.

Vijver, V. D., & Poortinga, Y. H. (1985). A Comment on McCauley and Colberg's Conception of Cross-Cultural Transportability of Tests. *Journal of Educational Measurement, 22*(2), 157-161.

Wang, W.-C. (2008). Assessment of Differential Item Functioning. *Journal of Applied Measurement, 9*(4).

Wu, A. D., & Ercikan, K. (2006). Using Multiple-Variable Matching to Identify Cultural Sources of Differential Item Functioning. *International Journal of Testing, 6*(3), 287 — 300.

Xie, Y., & 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.

Chapter 2: The Statistical DIF Detection Method

2.1 Data

Instrument: the PISA 2006 Science Scale

Assessing Scientific Literacy The PISA 2006 Science scale assesses the "scientific literacy" within a framework that situates three kinds of scientific competencies (i.e., identifying scientific issues, explaining phenomena scientifically, and using scientific evidence) within scientific contexts (i.e., life situations involving science and technology) in two science knowledge domains ("knowledge of science", referring to students' understanding of fundamental scientific concepts, and "knowledge about science", referring to students' understanding of scientific inquiry and scientific explanations). The assessed content areas include physical systems, living systems, earth and space systems, etc. In addition, for the first time the instrument includes students' attitudes toward science (i.e., "interest in science" and "support for scientific inquiry") as an important aspect of "scientific literacy". (These items were labeled as "attitude items" while items on student science knowledge and skills were called "cognitive times" in OECD official reports.) (OECD, 2006)

Test Format The PISA 2006 Science test arranged items in units based upon a common stimulus. Many different types of stimulus were used including passages with text, tables, graphs and diagrams. Each unit contained up to four items assessing students' scientific competencies and knowledge. In addition, about 60% of the science units contained one or two items designed to assess aspects of students' attitudes towards science. In the final version, there were 37 science units, comprising a total of 103 cognitive items at varying difficulty levels and 89 embedded attitudinal items, representing approximately 210 minutes of testing time.

To overcome the conflicting demands of limited individual testing time and a broad coverage of the science competency domains, students were assigned a subset of 192 science items. The items were presented to students in thirteen test booklets. According to a rotational design (OECD, 2006), each sampled student responded to one of the thirteen booklets by random assignment, and every item was answered by some students within each participating country.

Booklet Effect As it is extremely hard to make all the booklets equally difficult, it was expected that there would be some booklet influences on the estimated proficiency distributions.

The booklet effects are the amount that needs be added to the proficiencies of students who responded to each booklet. That is, a positive value indicates a booklet that was harder than the average while a negative value indicates a booklet that was easier

than the average. To correct the student scores for the booklet effects, the booklet parameters can be added to the students' achievement estimates.

However, booklet effect was not taken into consideration in this study because of two reasons. First, PISA reported that there is no significant booklet effect at international level (OECD, 2009a). The booklets were carefully designed and well balanced, and corrections were made to control for this effect. Secondly, the focus of this study does not lie in individual achievement estimates, but in item parameter estimates. Item parameter estimates that are obtained from our scaling procedure, with data from all participants on all items modelled together, are not influenced by a booklet effect.

Item Formats PISA 2006 was a pencil-and-paper assessment. Item formats employed with the science cognitive items were multiple-choice, short closed-constructed response, and open- (extended) constructed response. Multiple-choice items were either standard multiple choice with four responses from which students were required to select the best answer, or complex multiple choice items presenting several statements for each of which students were required to choose one of several possible responses (yes/no, true/false, etc.). (These statements were each counted as a dichotomously scored item in data calibrations.) Closed-constructed response items required students to construct a numeric response within very limited constraints, or only required a word or short phrase as the answer. Open-constructed response items required more extensive writing and frequently required some explanation or justification. For attitude items, students were required to express their level of agreement with two or three statements expressing either interest in science or support for science using four categories of responses. (Each statement was coded as a polytomous item for scaling purposes.)

Participants

PISA Sampling Design The target population of PISA is 15-year old students attending educational institutions, in grades 7 and higher. The age of the students participating in the test must be between 15 years and 3 (completed) months to 16 years and 2 (completed) months at the beginning of the testing window.

The sampling design used for the PISA was a two-stage stratified design. In the first stage, schools having 15-year-old students were sampled systematically from a comprehensive national list of all eligible schools with probabilities that were proportional to the estimated number of eligible 15-year-old students enrolled. This is referred to as systematic probability proportional to size (or PPS) sampling. The second-stage sampling units were students within sampled schools. Once schools were selected to be in the sample, a list of each sampled school's 15-year-old students was prepared. For each country a target cluster size (TCS) was set, this value was typically 35 although with agreement countries could use alternative values. From each list of students that

contained more than the TCS, the TCS students were selected with equal probability and for lists of fewer than the TCS, all students on the list were selected.

By applying the two-stage stratified sampling design, nationally representative samples were drawn. More than 400 000 students in 57 countries participated in PISA 2006. The final international sample represents 20 million 15-year-olds around the world (OECD, 2007). Besides, Mainland China, although not officially listed as a participating country, also took part in the assessment. It is the first time mainland China participated in PISA, so these results are particularly informative and valuable.

The selected samples for this study To answer the research questions of this study and investigate the measurement equivalence between the three comparison groups, the response data from U.S., English-speaking Canadian students, mainland Chinese, and Chinese Hong Kong students were put under scrutiny.

These four groups were chosen because they present opportunities to do three interesting comparisons: (1) U.S. and English-speaking Canada students use the same official language, have similar culture and similar curricula; (2) students from mainland China and Chinese Hong Kong speak similar languages (simplified Chinese vs. traditional Chinese as their written language, and mandarin vs. Cantonese as their spoken language, respectively), have similar culture but use different curricula; (3) whereas Chinese and U.S. students speak completely different languages, have very different culture and use different curricula.

5611 U.S. students, 17555 Canadian students, 4892 mainland Chinese students and 4645 Hong Kong students participated in the assessment. (Note that 22646 Canadian students took test, but 5091 students took the French version and 17555 students took the English version. For the purpose of this study, only responses to the English version were studied to exclude the translation factor in the analysis.)

The strength and limitations of the data The PISA 2006 science assessment data makes an ideal choice for exploring the cross-national DIF issue in international science tests. First, the sample size is large enough to warrant the application of any statistical DIF analyses methods. Secondly, the strict probability sampling procedure ensures the generalizibility of the research results. Third, as Science is the major domain in the 2006 assessment, the large number of science items covers a wide variety of competency domains, response formats and difficulty levels. The relatively large item pool allows for in-depth substantive analysis to locate the potential causes of DIF. Fourth, PISA 2006 has an extraordinarily broad geographical coverage, which offers the opportunity to investigate DIF between many cultures. In addition, the ongoing nature of PISA offers the possibility to conduct multi-year comparisons, which will in turn provide evidence for the consistency of the findings.

On the other hand, the data has its limitations as well. One major drawback is that the confidentiality of the item contents made it very difficult to carry out substantive analysis, let alone discuss the possible causes of the detected DIF in details. Even though permission was given for me to view the items, the time allowed for that was very limited. And "pseudo items" (items I made up that resemble the important features of the original ones) were used instead of the authentic ones for the purpose of discussing the possible causes of DIF with content experts, as well as reporting findings in this study.

2.2 Statistical DIF Detecting Methods

Existing DIF detection methods

To address the issue of cross-national comparability between the groups of students in this study, DIF analyses are needed. Researchers have been actively developing and refining psychometric procedures to detect the incomparability of items for several decades. Currently, there are many different statistical methods available for DIF detection. Some popular ones include the Mantel-Haenszel (MH) method (Holland & Thayer, 1988), the standardized *p*-difference index (Dorans & Holland, 1993), logistic regression (Swaminathan & Rogers, 1990), Raju's area measures (Cohen & Kim, 1993), SIBTEST (Shealy & Stout, 1993), and Rasch-based random coefficient multinomial logit model (RCMLM) for DIF detection (Meulders & Xie, 2004). Among these methods, the MH procedure, standardized *p*-difference index, and logistic regression method are based on observed scores, whereas Raju's area measures and Rasch-based logit models assume the unobserved latent variable underlying the assessed performance.

Each of these DIF detection methods has its own advantages and disadvantages (Millsap & Everson, 1993). The Mantel-Haenszel (MH) statistic is one of the most widely used methods in detecting item-level measurement bias, largely because it is conceptually simple, relatively easy to use, and provides a chi-square test significance. Moreover, besides a test of the null hypothesis, it also estimates the size of DIF in an item. However, the procedure also has several disadvantages. First, the test is not designed to detect non-uniform DIF (the direction of item bias is not the same for examinees with low and high total scores). The second problem concerns the adequacy of the total score as a substitute for the latent trait. Both theoretical studies and simulation studies (Meredith & Millsap, 1992; Millsap & Meredith, 1992) have shown that this procedure can falsely detect DIF when the responses are generated by complex models. In addition, several extensions of the MH procedure for polytomous data were found to have relatively large Type I error rate under no-DIF conditions.

Another highly rated DIF detection technique is the standardized *p*-difference approach. The results are very close to the MH procedure. However, the problems in the MH procedure will also affect the standardization procedure.

The logistic regression method is sensitive to both uniform and non-uniform DIF. It can be extended to multiple examinee groups and to polytomous items. However, this

procedure also shares the difficulties faced by the MH method and standardized *p*-difference approach. The use of the total score as a proxy of the latent trait will encounter problems when the responses follow complex IRT models. Relatively long tests can reduce the chances for problems to arise. In shorter tests, theory suggests that false indication of DIF will be encountered.

Modern item response theory assumes that there is a latent trait underlying the observed performance. Item response models create a continuum on which both student performance and item difficulty will be located and a probabilistic function links these two components. Many DIF detection methods have been developed that operate within an assumed item response model.

For example, Raju's area measures of bias express the difference between the reference and focal group as some function of the area between their item response functions (IRFs). The IRFs can be either one-, two- or three- parameter logistic functions. Many choices for the function are possible: signed or unsigned differences, bounded or unbounded interval on the latent trait scale. One difficulty of applying this DIF detection method is that the choices between bounded and unbounded area measures, and between signed and unsigned area indexes remain unclear. Another long recognized problem is that the standard errors (SEs) of the measures are not calculated, making it difficult to evaluate the statistical significance of any differences found. Wald chi-square statistics (Lord, 1980) permit the use of SEs for statistical significance test. The null hypothesis of identical IRFs is tested under an assumed parametric model. One criticism is that the null hypothesis may be rejected even the unsigned area is very small when sample size is large.

The SIBTEST method resembles the standardized *p*-difference index (Shealy & Stout, 1993). The procedure begins by identifying a subset of items as the "valid subtest", which is a group of items believed to be measuring only the target trait. This test is more sensitive to unidirectional bias, that is, the reference group is expected to score as well or better than the focal group.

The Rasch-based RCMLM method is a logistic function that estimated the group-by-item interaction and tests the null hypothesis that the interaction is zero. It can be easily extended for polytomous items and multiple comparison groups. One major criticism of this procedure is the sufficient statistic assumption, that is, the total score sufficiently represents the latent variable. Sufficiency breaks down when data is generated by 2- or 3-parameter IRT models. Problems may arise using this model to detect bias in such data, especially when sample size is relatively small.

Choice of the DIF detection method for this study

As each of the reviewed DIF detection methods has its own strengths and drawbacks, the choice of which procedure to apply is not an obvious one. However, the Rasch-based RCMLM procedures seem to be the most appropriate for this study for

several reasons. First, PISA employs Rasch model to estimate student ability and item difficulty, and create the overall PISA literacy scale. Hence, the items used in the PISA studies have been checked for adherence to the Rasch model.

Several important principles underlie the Rasch model: first and most importantly, only one factor, the distance on the Rasch continuum between the student ability and the item difficulty, influences the probability of success. When the student ability is equal to the item difficulty, the probability of success will always equal 0.50, regardless of the student ability and item difficulty locations on the continuum. Secondly, the relative difficulty of an item results from the comparison of that item with all other items. It is independent of the student abilities. And similarly, the examinee ability estimates can be interpreted independently of the item difficulties. As a result, the Rasch model creates a relative scale, under which only one reference point needs to be defined. The most common reference point consists of centering the item difficulties on zero. But other arbitrary reference points can be used, like centering the student's abilities on zero. Because of these properties, the Rasch model was chosen for the scaling of PISA data. The Rasch-based RCMLM procedures yield results that are consistent and comparable to those in PISA official reports and other related papers.

A second reason why the RCMLM procedures are chosen for this study is that it is technically more desirable than other IRT-based methods. A previous analysis (Huang, 2005) has found that this approach yields very similar results as the area measures method, but it utilizes information from all participants simultaneously instead of information from only one group each time in the estimation procedure, thus, the standard errors of the item difficulty estimates for different comparison groups can be compared directly.

In addition, this RCMLM approach can easily be extended to analyze polytomous items, and be adapted to a multidimensional framework. There is an existing software, namely, ConQuest (M. Wu, Adams, & Wilson, 1998), that can be used to implement these procedures. And finally, as the PISA 2006 Science assessment contains 192 items, the sufficient statistic assumption is less of a concern (Millsap & Everson, 1993).

The unidimensional RCMLM

For the reasons stated above, the RCMLM DIF detection procedure is selected to carry out the statistical DIF analysis in this study.

The RCMLM is a generalized Rasch model that integrates many other kinds of Rasch models (Adams, Wilson, & Wang, 1997), such as the simple logistic model (Rasch, 1980), the partial credit and the rating scale model (Wright & Masters, 1982). When the items are dichotomously scored, the RCMLM is adjusted to be the simple logistic model (the Rasch model), and when the items are polytomously scored, the RCMLM is adjusted to be the partial credit or rating scale model.

The RCMLM describes the items by a set of fixed parameters $\boldsymbol{\xi}$, and the examinee latent ability by a random variable θ (a scalar). Assume that N examinees are indexed as n = 1, … , N; and I items are indexed as i = 1, … ,I with each item admitting $K_i + 1$ response categories, indexed as k = 0,1,...,K_i. A vector of valued random variable $\mathbf{X}_{ni} = (X_{ni1}, X_{ni2},..., X_{niKi})^T$ indicates the $K_i + 1$ responses to item *i* from examinee *n*, where

$X_{ij} = 1$ if response to item i is in category j,
0 otherwise.

The latent ability that is being measured is denoted as θ_n for examinee *n*. The vector of item parameters for item *i* $\boldsymbol{\xi}_i = (\xi_{i1}, \xi_{i2}, \dots , \xi_{ip})$ consists of *p* parameters. Linear combinations of these parameters are used in the response probability model to describe the empirical characteristics of the response categories of each item. A *design matrix* is used to define the linear combinations of ξs. Design vectors $\mathbf{a}_{ij}$ (i = 1, … , I ; j = 0, 1, … , K_i), each of length p, can be collected for all items to form the design matrix $\mathbf{A} = (\mathbf{a}_{11}, \mathbf{a}_{12}, \dots , \mathbf{a}_{1Ki}, \mathbf{a}_{21}, \dots , \mathbf{a}_{2K2}, \dots , \mathbf{a}_{IKi})$.

The model also introduces a scoring function that specifies the scores assigned to each response category of each item. The vector $\mathbf{b}_i = (b_{i0}, b_{i1}, \dots, b_{ik})$ gives the performance level of an observed response in category j item i, $b_{ij}=j$ if the response falls into that category. The **b** vectors can again be collected into a *scoring matrix* $\mathbf{B} = (\mathbf{b}_1, \mathbf{b}_2, \dots , \mathbf{b}_i)$ for the entire test.

The probability of a response in category *j* of item *i* for examinee *n* is modeled as

$$\Pr\left(X_{nij}=1;\mathbf{A},\mathbf{B},\boldsymbol{\xi}_i\middle|\theta_n\right)=\frac{\exp\left(b_{ij}\theta_n-\mathbf{a}'_{ij}\boldsymbol{\xi}_i\right)}{\sum_{k=0}^{K_i}\exp\left(b_{ik}\theta_n-\mathbf{a}'_{ik}\boldsymbol{\xi}_i\right)}. \qquad \text{(Equation 2.1)}$$

The scoring function and the design matrix introduces great flexibility to the RCMLM. The former allows users to specify individual item weight (where priori weights are decided based on some theoretical or practical reasons) and the latter allows users to specify item parameters in a linear form. For example, assigning equal weight to all items in the scoring function and specifying only one item difficulty parameter for each item in the design matrix leads to the simple logistic (Rasch) model for dichotomous response data.

Within the framework of the RCMLM, the DIF detection model for person *n* to score 1 on a dichotomous item *i* can be expressed as

$$\Pr\left(X_{ni}=1\middle|\theta_n,G_n\right)=\frac{\exp\left[\theta_n-\left(\xi_i+\gamma_i G_n\right)\right]}{1+\exp\left[\theta_n-\left(\xi_i+\gamma_i G_n\right)\right]}, \qquad \text{(Equation 2.2)}$$

where γ_i is the item DIF parameter, with $G_n=0$ if the examinee belongs to the reference group and $G_n=1$ for the focal group. The examinee ability estimate θ_n consists of two parts: $\theta_n=\alpha_{Gn}+e_n$, α_{Gn} being the group mean ability and e_n being the individual difference from the group mean. Note that for the reference group where $G_n=0$, the model becomes simple Rasch model:

$$\Pr\left(X_{ni}=1\middle|\theta_n,G_n\right)=\frac{\exp\left[\theta_n-\xi_i\right]}{1+\exp\left[\theta_n-\xi_i\right]}. \qquad \text{(Equation 2.3)}$$

When G is coded this way, γ_i can be interpreted as the difference in the item difficulty between the focal and the reference group. The logits of success is θ_n- ξ_i for the reference group, ξ_i being the difficulty of item *i*. For the focal group, the logits of success is expressed as θ_n-(ξ_i+γ_i), γ_i being the group-by-item interaction effect, which is the difference between group-specific item locations. The absence of DIF is modeled by the null hypothesis γ_i=0. The Wald test can be used to test whether individual items have statistically significant DIF.

This model can be easily extended to polytomous items. The two most commonly used models for polytomous response data in the Rasch family are the partial credit model (PCM) and the rating scale model (RSM). For PISA 2006 science items, the PCM is applied. The RSM is not appropriate here because the number of response categories differs across items, and we cannot assume a common value of "threshold" between successive response categories for all items.

The PCM includes *K-1* ξ parameters for each K-category item. Wright and Masters (1982) described these parameters as step-difficulties (ξ_{ij}), viewing the completion of an item involving *K-1* steps. An examinee's score on an item denotes the number of steps completed. The RCMLM DIF model for person *n* to score *j* on polytomous item *i* can be expressed as

$$\Pr\left(X_{nij}=1\middle|\theta_n,G_n\right)=\frac{\exp\sum_{k=1}^{j}\left[b_{ik}\theta_n-\left(\xi_{ik}+\gamma_i G_n\right)\right]}{\sum_{\mathbf{j=1}}^{\mathbf{K_i}}\exp\sum_{k=1}^{j}\left[b_{ik}\theta_n-\left(\xi_{ik}+\gamma_i G_n\right)\right]}. \qquad \text{(Equation 2.4)}$$

X_{nij} = 1 if response to item i is in category j,
0 otherwise.

Again, for the reference group where G=0, the model becomes the PCM:

$$\Pr\left(X_{nij}=1\middle|\theta_n,G_n\right)=\frac{\exp\sum_{k=1}^{j}\left[b_{ik}\theta_n-\xi_{ik}\right]}{\sum_{\mathbf{j=1}}^{\mathbf{K_i}}\exp\sum_{k=1}^{j}\left[b_{ik}\theta_n-\xi_{ik}\right]}. \qquad \text{(Equation 2.5)}$$

Similarly, the logits of success is θ_n-ξ_{ij} for the reference group, and θ_n-(ξ_{ij}+γ_i) for the focal group. The interpretation of γ_i remains the same as in the dichotomous model. The Wald Chi-square test can also be used to test whether an item has statistically significant DIF.

Note that in this model we assume the group-by-item interaction is at item level, that is, there is only one DIF parameter (γ_i) for each item. It is also possible to estimate step-specific interaction effect. When DIF occurs only in the overall item difficulty, one group always has a higher expected score than the other group throughout the latent trait level on that item. Whereas when DIF occurs at the threshold level, the magnitude and direction of DIF may change across the range of latent trait level. Substantive analysis on

step-specific DIF effects is too complicated to implement in this study. So here, we assume DIF only at item level.

The Multidimensional RCMLM

The DIF detection procedures discussed thus far are based on the assumption of a unidimensional latent trait underlying test performance. However, researchers have found that when the percentage of DIF bias was high (20% or higher), and especially with large magnitude, traditional unidimensional approach cannot identify DIF precisely (Miller & Oshima, 1992).

Some suggest that item bias might potentially be introduced by a secondary dimension to the unidimensional estimates of the person and item parameters (Millsap & Everson, 1993). The initial unidimensional parameter estimates may be a weighted composite of two or more traits. At a minimum, the unidimensional estimates are affected by some differential variability on the nuisance dimension.

For PISA 2006 Science assessment data, the unidimensionality assumption may be at risk. At the international level, the concept of "scientific literacy" is multifaceted (Schwab, 2007). Science educators view scientific literacy as a multidimensional construct and have established a body of literature describing its varied components (Laugksch, 2000). Yet the measurement of scientific literacy has been focusing on measuring this construct either as a single dimension or as multiple constructs separately (i.e., content knowledge, science ability, nature of science, and science and society). The vast majority of assessments developed by science educators are unidimensional in nature and target specific science content knowledge (Laugksh, 2000). PISA 2006 Science scale offers a unique opportunity to evaluate the dimensionality property of "scientific literacy".

Two three-dimensional models were proposed to scale the PISA 2006 Science assessment data (OECD, 2009a). The first model is made up of one science cognitive dimension and two attitudinal dimensions (i.e., "interest in learning science" and "support for scientific inquiry"). The correlations between the three dimensions for all participants are shown in Table 1 below (OECD, 2009a).

A second model consists of three science competency dimensions, namely, (1) explaining phenomena scientifically, (2) identifying scientific issues and (3) using scientific evidence. Table 2 shows the correlations between the three competency dimensions for all participating countries (OECD, 2009a).

Table 1: Correlations between Science Cognitive and Attitude Dimensions

	Cognitive Science Items	Interest in Science	Support for Scientific Inquiry
Cognitive Science Items	1		
Interest in Science	0.06	1	
Support for Scientific Inquiry	0.60	0.25	1

Table 2: Correlations between Three Science Competency Scales

	Explaining phenomena scientifically	Identifying scientific issues	Using scientific evidence
Explaining phenomena scientifically	1		
Identifying scientific issues	0.90	1	
Using scientific evidence	0.91	0.93	1

From the two tables, we can see that the correlations between the three competency scales are very high with values larger than 0.9; whereas the correlations between cognitive items and the two interest scales are much lower, with the correlation between cognitive items and "interest in science" as low as 0.06. So it seems safer to assume unidimensionality among the three competency scales. But the alarmingly low correlations between the cognitive items and the "interest" items do arouse concerns about the unidimensionality assumption. Hence, the three-dimensional model consisting one cognitive dimension and two interest dimensions is called for especially when a large amount of items were found to display DIF in the unidimensional analysis. The three-dimensional model may reduce DIF to some extent, or, it can real which dimension most detected DIF item reside.

In practice, the multidimensional procedure has not been widely applied. It is only recently that models within IRT have been proposed for DIF investigations under a multidimensional framework (Einarsdóttir & Rounds, 2009; Stark, Chernyshenko and Drasgow, 2006; Walker, Zhang & Surber, 2008). The multidimensional version of the RCMLM, called the multidimensional random coefficient multinomial logit model (MRCMLM) (Adams, et al., 1997), was adopted to examine student performance under this circumstance. The multidimensional form of the model assumes that a set of D traits underly the individuals' responses. The D latent traits define a D-dimensional latent space. The vector $\boldsymbol{\theta}_n = (\theta_{n1}, \theta_{n2}, \ldots, \theta_{nD})'$, represents an individual's position in the D-dimensional latent space.

It models the probability of a response in category j of item i for a person with latent trait $\boldsymbol{\theta}_n$ as:

$$\Pr\left(X_{nij}=1;\mathbf{A},\mathbf{B},\xi_i|\theta_n,G_n\right)=\frac{\exp\sum_{k=1}^{j}\left[\mathbf{b}_{ik}'\theta_n-\left(\mathbf{a}_{ik}'\xi_i+\gamma_i G_n\right)\right]}{\sum_{j=1}^{K_i}\exp\sum_{k=1}^{j}\left[\mathbf{b}_{ik}'\theta_n-\left(\mathbf{a}_{ik}'\xi_i+\gamma_i G_n\right)\right]}. \quad \text{(Equation 2.6)}$$

X_{ij} is defined the same as in the RCMLM, which is a person's response on item i, category j, $X_{ij}=1$ if the response falls into that category, and $X_{ik}=0$ if not. $\boldsymbol{\theta}_n$ is a D x 1 column vector for a person's D traits. Within each dimension, θ_{nd} can be viewed as a latent regression of the mean ability α_{Gnd} ($\theta_{nd}=\alpha_{Gnd}+e_{nd}$). ξ_i is a vector of item parameters across dimensions. $\mathbf{b}_{ij}$ is a column vector of a person's score on item i category j in each of the D dimensions, $\mathbf{b}_{ij}=(b_{ij1}, b_{ij2}, \ldots, b_{ijD})^T$. The vector can again collected into the scoring sub-matrix across D dimensions for item i, $\mathbf{B}_i=(\mathbf{b}_{i1}, \mathbf{b}_{i2}, \ldots, \mathbf{b}_{iD})^T$, and then into a scoring matrix $\mathbf{B}=(\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_I)$. $\mathbf{a}_{ij}$ is still the design vector of item *i,* which specifies the linear combination of the item's difficulty parameters for each response category across dimensions. γ_i and G are defined the same as in the RCML model, with γ_i being the DIF parameter and G indicating the group identity. A Chi-square statistic can also be used to test the significance of the difference in item locations.

DIF effect size

As mentioned earlier in this chapter, one major shortcoming about using the Chi-square test to examine the severity of DIF is that no matter how minute an item's amount of DIF, it will be detected as statistically significant as long as the sample size is sufficiently large. As the sample sizes are quite large in this study, it is highly likely that a trivial DIF would be identified as statistically significant. Thus, in addition to testing the statistical significance, it is also necessary to investigate the magnitude of DIF using an approach that is not dependent on sample size. Within the family of Rasch models, the magnitude of DIF is actually an effect size measure of the between-group difference in item parameters.

A widely used approach for categorizing the level of differential item functioning (DIF) in dichotomous items is the scheme proposed by Educational Testing Service (ETS) (Penfield, 2007). An item is said to display negligible DIF (Class A DIF) if the absolute value of the group difference is smaller than 0.426, intermediate DIF if group difference is smaller than 0.638 (Class B DIF), and large DIF (Class C DIF) if the difference is larger than 0.638. Only those items with intermediate to large DIF effect sizes would be identified as DIF items.

While the ETS classification scheme is widely used for dichotomous items, an analogous classification scheme has been proposed by Penfield (2007), which permits a consistent classification basis for dichotomous and polytomous items. In this study, the estimator of DIF effect size for polytomous items holds a theoretical equivalence with the one used for dichotomous items—the group*item interaction parameter. Similarly, a parallel scheme for classifying DIF in polytomous items as negligible (A), moderate (B), and large (C) is: negligible (A) if the absolute value of the group difference is smaller

than 0.426, intermediate (B) if group difference is smaller than 0.638, and large DIF (C) if the difference is larger than 0.638.

These industrial DIF classification criterion have been linked to ConQuest results (Paek, 2002). The same critical values hold for ConQuest results: an item is said to display negligible DIF (Class A) if the absolute value of twice of the ConQuest estimate of item-by-group interaction ($2|\gamma|_i < 0.426$), intermediate DIF (Class B) if $0.426 \leq 2|\gamma|_i < 0.638$, and large DIF (Class C) if $0.638 \leq 2|\gamma|_i$ (Paek, 2002).

2.3 Systematic DIF

The identification of isolated DIF items is not necessarily seen as a serious problem (AERA, APA, & NCME, 1999), but systematic and consistent findings of DIF are definitely problematic. When an assessment has a large number of items with large DIF effects, it is qualitatively different for different participating groups (e.g., different language users) so that the assumption of functional equivalence can no longer hold over the groups. Hence, in addition to examining the practical significance of DIF at the item level, it is also necessary to ascertain differential functioning at the test level (*DTF*).

Empirical studies showed that the percentage of DIF items can range from quite small (1.5%) to overwhelmingly large (64%). Simulation studies (Ronald K. Hambleton & Rogers, 1989; N. S. Raju, 1989) consider it a small amount of DIF when a test contains less than 10% DIF items, a medium amount of DIF when a test contains 10 to 30% DIF items and a large amount of DIF when the percentage of DIF items exceeds 30%. It is obvious that when the percentage of DIF items exceeds 10%, closer attention should be paid.

There are two widely used methods to evaluate DIF at test level. One is to compare the test characteristic curves for different groups of test takers, which is referred to as the assessment of differential test functioning (Raju, van der Linden, & Fleer, 1995). If the test expected score curves for different groups are far apart, then the DIF items are practically significant at test level. On the contrary, if the curves nearly overlap, then the DIF items are not practically significant, even though the test may contain a high percentage of DIF items that are statistically significant at the item level.

The other method is to compare person measures obtained from a model in which DIF items are excluded with those obtained from another model in which DIF items are not excluded. If the person measures obtained from these two models are very different, then the inclusion of DIF items substantially affects person measures. This method is selected in this study because the severity of DIF at the test level does not rely on subjective judgment. A simple t-test can easily show whether the person ability estimates obtained from the two models are significantly different, and the magnitude of the difference can be depicted by the effect size indicator.

When a test contains more than 10% DIF items and the t-test suggests the practical significance at test level, multidimensional analyses is followed and detailed content analyses on problematic items is conducted.

2.4 Data Calibration Software and Procedure

The statistical package SAS (SAS Institute Inc, 2000) was used for data cleaning, merging, recoding, and getting descriptive statistics as well as performing t-tests. The computer program *ConQuest* (Wu, Adams, and Wilson, 1998) is used for both unidimensional and multidimensional DIF detection analysis. For each pair of the comparison groups, the data is first analyzed under the unidimensional framework. *ConQuest* estimates the overall item difficulties and the difference between the means of each group. It will also provides an estimate of the "item*group" interaction term for each item, which is the γ_i term in the equations 2.2, 2.4 and 2.6, and the standard errors for the item-by-group interaction. The statistical significance of this estimate is indicated by the ratio of the estimate to its standard error of measurement. The ETS effect size classification rules are applied. Items with statistically significant γ_i will be classified into three categories. If the number of class B and C DIF items exceeds 10% (medium to high percentage of DIF items, Hambleton & Rogers, 1989; Raju, 1989), a multidimensional analysis is followed. Otherwise, validity equivalence between the two groups of students can be safely claimed.

Under the multidimensional framework, the γ_i term is again examined by comparing its absolute value to the ETS effect size classification rules. If the number of class B and C DIF items reduced significantly (to less than 10%), the data is advised to be scaled under multidimensional framework. Accordingly, discussions on cross-country comparisons should be based on the multidimensional calibration results as well. On the other hand, if the percentage of DIF items is still not negligible (>10%), a t-test is performed to examine whether the DIF has practical significance at test level. In addition, a detailed substantive analysis seeking potential causes of the detected DIF problem is required.

The unidimensional→multidimensional (if necessary)→content-analysis (if necessary) cycle is repeated three times for the three pairs of comparison groups.

References

Adams, R. J., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficient multinomial logit model. *Applied Psychological Measurement, 21*(1), 1-23.

AERA, APA, & NCME (1999). *Standards for Educational and Psychological Testing. .* Washington, DC: American Educational Research Association.

Cohen, A. S., & Kim, S.-H. (1993). A Comparison of Lord's Chi Square and Raju's Area Measures in Detection of DIF. *Applied Measurement in Education, 17*(1), 39-52.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & W. Howard (Eds.), *Differential Item Functioning* (pp. 35-66). Hillsdale: NJ: Lawrence Erlbaum.

Einarsdóttir, S., & Rounds, J. (2009). Gender bias and construct validity in vocational interest measurement: Differential item functioning in the Strong Interest Inventory. *Journal of Vocational Behavior, 74*(3), 295-307.

Hambleton, R. K., & Rogers, H. J. (1989). Detecting Potentially Biased Test Items: Comparison of IRT Area and Mantel-Haenszel Methods. *Applied Measurement in Education, 2*(4), 313-334.

Holland, P. W., & Thayer, D. T. (1988). Differential item funtioning and Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp. 129-145). Hillsdale:NJ: Lawrence Erlbaum.

Huang, X. (2005). *Validity Equivalence between the Chinese and English Versions of the IEA Child Cognitive Developmental Status Test* Paper presented at the International Objective Measurement Workshop.

Laugksch, R. C. (2000). Scientific literacy: A conceptual overview. *Science Education, 84*(1), 71-94.

Lord, F. M. (1980). *Applications of Item Response Thoery to Practical Testing Problems*. Hillsdale, NJ: Erlbaum.

Meredith, W., & Millsap, R. (1992). On the Misuse of Manifest Variables in the Detection of Measurement Bias. *Psychometrika, 57*(2), 289-311.

Meulders, M., & Xie, Y. (2004). Person-by-item Predictions. In P. Boeck & M. Wilson (Eds.), *Explanatory Item Response Models* (pp. 213-240). New York: Springer.

Miller, M. D., & Oshima, T. C. (1992). Effect of Sample Size, Number of Biased Items, and Magnitude of Bias on a Two-Stage Item Bias Estimation Method. *Applied Psychological Measurement, 16*(4), 381-388.

Millsap, R., & Everson, H. T. (1993). Methodology Review: Statistical Approaches for Assessing Measurement Bias. *Applied Psychological Measurement,, 17*(4), 297-334.

Millsap, R., & Meredith, W. (1992). nferential Conditions in the Statistical Detection of Measurement Bias. *Applied Psychological Measurement, 16*(4), 389-402.

OECD (2006). *Assessing Scientific, Reading and Mathematical Literacy: a Framework for PISA 2006.*

OECD (2007). *PISA 2006: Science Competencies for Tomorrow's World* (Vol. 1).

OECD (2009). *PISA 2006 Techinical Report.*

Paek, I. (2002). *Investigations of differential item functioning: comparisons among approaches, and extension to a multidimensional context.*, University of California, Berkeley, Berkeley.

Penfield, R. D. (2007). An Approach for Categorizing DIF in Polytomous Items. *Applied Measurement in Education, 20*(3), 335-355.

Raju, N. S. (1989). An Empirical Assessment of the Mantel-Haenszel Statistic for Studying Differential Item Performance. *Applied Measurement in Education, 2*(1), 1-13.

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differ- ential functioning of items and tests. *Applied Psychological Measurement, 19*, 353-368.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: The University of Chicago Press.

SAS Institute Inc (2000). SAS Version 9.1. Cary, NC, USA.

Shealy, R., & Stout, W. (1993). An item response theory model for test bias and differential item functioning. In P. W. Holland & W. Howard (Eds.), *Differential item functioning*. Hillsdale: NJ: Lawrence Erlbaum.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology, 91*(6), 1292-1306.

Schwab, C. J. (2007). *What Can We Learn from PISA? Investigating PISA's Approach to Scientific Literacy.* University of California, Berkeley, Berkeley.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 361-370.

Walker, C. M., Zhang, B., & Surber, J. (2008). Using a Multidimensional Differential Item Functioning Framework to Determine if Reading Ability Affects Student Performance in Mathematics. *Applied Measurement in Education, 21*(2), 162-181.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.

Wu, M., Adams, R. J., & Wilson (1998). ConQuest. Hawthorn, Australia: ACER Press.

Chapter 3: Investigating the Causes of DIF

3.1 Review of Potential Causes of DIF in International Assessments

Just showing whether an item has statistically substantial DIF is not sufficient. Especially when a considerably large number of items in a test exhibit DIF, we have reason to doubt whether the test measures the same psychological constructs in different sub-populations, and whether the measured variables are those that are intended or other unintended or unwanted variables. Further efforts should be exerted to investigate the underlying causes of the detected DIF effects.

Previous research has identified many possible sources of DIF in international assessments, such as test translation (Allalouf, 2000; Ercikan, 1998; Price & Oshima, 1998; Sireci & Swaminathan, 1996; Wolf, 1998), differential curriculum coverage (Beaton, 1998; Schmidt, Jakwerth, & McKnight, 1998; Westbury, 1993), differential impact of item format (Allalouf et al, 1999; Gierl & Khaliq, 2001), differential cultural appropriateness of item content (Vijver & Poortinga, 1997; Wu & Ercikan, 2006), speediness and other physical test conditions (Vijver & Poortinga, 1997), etc. Among these factors, test translation, curriculum differences and cultural differences are most frequently discussed.

Language Difference

Literature on empirical comparative research refers to test translation issues as one of the most frequent problems in cross-cultural assessments (Allalouf, 2000; Ercikan, 1998; Hambleton & Kanjee, 1995; Hambleton & Patsula, 1999; Price & Oshima, 1998; Sireci & Swaminathan, 1996; Wolf, 1998). Ideally, a single common version of a test would be used for international assessments. However, since international assessments are always administered in different countries, it is necessary to translate the tests into the languages of those countries. Test translation, as a potential source of DIF, may alter the items in terms of the meaning, the connotations, and the degree of difficulty of the key vocabulary, as well as the general style of the passages. The degree and manner in which item features are changed during the test translation process will determine how well the equivalence of the items is maintained.

There are several basic differences in languages that cause problems in test translation. First, the vocabulary difficulty may vary after translation due to differential frequency of word use in different languages or cultures. Secondly, problems occur when grammatical forms either do not have equivalents in different languages, or else have more of them in one language than the other. Syntactical style is extremely difficult to carry over from one language to another, resulting in differential length or complexity of sentences. In addition, it is hard to convey the same contextual meaning of vocabulary across languages (Ercikan, 1998).

One way to avoid translation-related problems is to use *forward translation* and *back translation* when creating multiple language versions of international assessments (Hambleton & Patsula, 2000). Forward translation translates the test from its source language to the target language, while back translation translates the new version back to the source language (usually by a different translator). The back-translated version can then be compared to the original one, and discrepancies can thus be identified and addressed. There are a few variations of this method. For example, TIMMS uses multiple forward translations by more than one translator and compares the translated versions (Wolf, 1998), and PISA translates the test from two different source languages and compares the translated versions (OECD, 2009a).

Despite the precautions taken at the test translation process, DIF between different languages is a big concern in international assessments (Allalouf, et al., 1999; Ercikan, 1998; Ercikan & Koh, 2005). Hence, it is an important direction for research to explore the causes of DIF in the PISA 2006 Science assessment, which is an important example of international testing.

Curriculum coverage

Differential curriculum coverage in different countries poses special challenges in the development of valid cross-national measures because the curriculum differences can result in varying degrees of student exposure to the content and processes required to answer the items correctly. The differential curriculum coverage inevitably leads to differential response patterns and difficulty levels, independent of any problems due to test translation. If a test has more items that are appropriate for some countries' curriculum than others, it brings into question the adequacy of the test for cross-national comparisons.

Balancing the coverage of content and process across different educational systems is very difficult. Students in different countries, and in different streams (or tracks) within countries, are exposed to different subject matters and teaching methods. Different countries also place different emphases on different topics. In addition, the order in which topics are introduced and subsequent instructions provided to students are different as well. Limiting a test to topics covered by all streams in all countries would result in a test so narrow as to be trivial. On the other hand, covering all possible topics would result in a test too long to be realistic.

In order to ensure the technical validity of findings from cross-national comparisons, researchers introduced the Opportunity to Learn (*OTL*) concept (McDonnell, 1995). OTL data is traditionally collected by polling teachers of the tested students on the opportunities their students have had to learn the content and processes needed to answer test items correctly. It is usually done in a survey form and requires item-by-item rating of all items. Sometimes, the students themselves are surveyed (Schmidt, 1998).

One critique about teacher OTL ratings is that the ratings are likely to be confounded by teacher perceptions of student opportunities, teacher reactions to the test form, and teacher perceptions of the likely student achievement, etc. However, researchers have shown that these ratings are significant explanatory variables for student performance (Berliner, 1993; Westbury, 1993). For example, Westbury (1993) found that differences between the scores of American and Japanese students on the Second International Mathematics Study (SIMS) decreased when controlling for curriculum through OTL data.

Another similar approach to assess the fairness of curriculum coverage is described as the Test-Curriculum Matching Analysis (TCMA) (Beaton, 1998). The TCMA collects national judgments on the appropriateness of the content of each test item. Each participating country is asked to indicate whether the content of each question was covered in that country's curriculum (a rating of 1) or not (a rating of 0). From these ratings of 1's and 0's, every country's percentage correct score was calculated and compared with those of other countries.

Research on the impact of curriculum on international assessment's fairness shows somewhat mixed results (Beaton, et al., 1996; Berliner, 1993; Schmidt, Jakwerth, & McKnight, 1998). Burstein (1993) found that countries performed better on ''custom'' tests developed to match their curricula and that the score ranges changed across the various customized tests. However, Beaton (1996) reported that allowing countries to select the items on which they are scored would have little effect on their international standings in TIMMS. But the sub-area rankings are different from the overall rankings. For example, Schmidt (1998) found that countries' relative standings could be unstable over different subareas or domains in TIMMS. The limited variability across total scores of those curriculum-based "customized" tests might partially be attributed to an inadequate item pool. Or, it could be because that not all topics were measured in depth.

Nevertheless, when a large amount of DIF is detected in an international assessment, differential curriculum coverage is undoubtedly an important place to look for reasons. OTL, TCMA or similar approaches can be applied to examine the impact of curriculum. It is unclear whether the variations of curricula within countries affect these results, although it is reasonable to assume it has some effects.

Cultural difference

Besides test translation and curriculum coverage, cultural difference is believed to be another major contributor to DIF (Vijver & Poortinga, 1997). However, research investigating the role of students' cultural background has been limited. Cultural sources of bias are often neglected because these differences are hard to investigate methodologically and their inferential power is often restricted by other confounding factors.

Existing studies focus on three important aspects of cultural difference: differential content familiarity, different social desirability, and difference in response style. First, cultural differences can influence the familiarity of the content of items. Typically, those items are judged to contain substantive content that may be more familiar to examinees in one culture than in the other. Or, they may include social practices that are unique in one country. For example, students in equatorial countries may have more difficulty in answering science items about seasonal changes than students from countries in which seasons vary. Thus, contextualizing items to make them more realistic (which is more common in some tests than others) is likely to introduce differences in complexity attributable to national variances.

Another important reason for the occurrence of differential cultural appropriateness is the issue of social desirability (Buckley, 2009). Norms about appropriate conduct differ across cultural groups, and the social desirability expressed in assessment varies accordingly. For example, Cheung (1996) observed differential responses in self-reported rates of depression among Chinese and American adolescents and concluded that these differences may be due to different attitudes about health as well as different levels of stress relating to academics between these two cultures (Cheung, 1996).

In addition, systematic differences in response style across nations or cultures also jeopardize the validity of inter-group comparisons. This is especially true for attitude survey items.

There are four typical response styles, namely the acquiescence response style (*ARS*), which is a tendency to agree with items regardless of actual attitude; the disacquiescence response style (*DARS*), a tendency to disagree regardless of their actual attitude; the extremity response style (*ERS*), which is a tendency to choose the endpoints of an item's scale regardless of the actual attitude; and the non-contingent responding (*NCR*), which is a term used to describe random or careless response to items (Buckley, 2009).

There is much empirical evidence of systematic differences between cultures in those response styles (Bachman & O'Malley, 1984; Chen, Lee, & Stevenson, 1995; Marin, Gamba, & Marin, 1992; Watkins & Cheung, 1995). For example, Chen, Lee, and Stevenson (1995) reported that Chinese and Japanese secondary students were more likely to use the midpoint of a seven-point Likert-type item, while U.S. students exhibited a greater tendency toward ERS than their Asian counterparts. Watkins and Cheung (1995) examined response styles of high school students from five countries and reported substantial variation in ERS and NCR on academic self-esteem items. Marin, Gamba, and Marin (1992) compared Hispanics to non-Hispanic Whites and found a greater incidence of both ERS and ARS among the Hispanic population, particularly the less educated and less acculturated. Bachman and O'Malley (1984) found similar results comparing black with white respondents.

These findings, particularly the cross-national research on secondary school populations, suggest that heterogeneity in response styles could be a potential source of bias in PISA.

3.2 Potential Causes of DIF in PISA 2006 Science Assessment

As the aim of PISA is to develop a scale that provides reliable and fully comparable information for all participating parties, the test development team implemented many strict procedures in order to achieve this goal. However, despite all the precautions, there is evidence that test translation, differential curriculum coverage and cultural difference may still cause functional inequivalence in the PISA 2006 Science scale.

Language Difference

Targeted Precautions To minimize the impact of language difference, PISA's test translation procedure includes the follow steps (OECD, 2009a):

- Development of two source versions of the instruments (in English and French);
- Double translation design;
- Preparation of detailed instructions for the translation of the instruments;
- Training of national staff in charge of the translation/adaptation of the instruments;
- Verification of the national versions by international verifiers.

Two source languages were used because using one single reference language is likely to give undue importance to the formal characteristics of that language. The lexical and syntactic features, stylistic conventions and the typical patterns the source language uses to organize ideas will have a greater impact on the target language versions than desirable (Grisay, 2003).

After the two source versions were developed, a "double translation" procedure was implemented. As reviewed in the previous section, back translation has long been the most frequently used procedure to check the linguistic equivalence of multiple language versions of international surveys. Double translation, which requires two independent translations from the source language(s) and reconciliation by a third person, has not been widely applied.

The double translation design offers two significant advantages in comparison with the back translation design: First, equivalence of the source and target versions is obtained by using three different people (two translators and a reconciler) who all work on both the source and the target versions. In a back translation design, by contrast, the first translator is the only one to simultaneously use both the source and target versions. Secondly, discrepancies are recorded directly in the target language instead of in the source language, as would be the case in the back translation design.

In addition, PISA provided test translation/adaptation guidelines to further ensure the comparability between different language versions. The guidelines include recommendations to avoid common translation traps and instructions on how to adapt the test material to the national context.

Furthermore, PISA also engaged an independent team of experts to verify each language version against the English and French source versions. The expert verifiers identified various errors, such as mistranslations, awkward expressions, incorrect terminology, poor rendering of graphics or layout, errors in numerical data, grammar and spelling errors (OECD, 2009a).

Remaining Problems Although PISA exerted a great amount of effort in test translation/adaptation, there is evidence that language effect may still be a potential cause of item bias. Specifically, correlations between the item parameter estimates betweem countries within a particular language as well as the correlations between these item parameter estimates and the international item parameter estimates were compared. If a language effect is suspected, then the within language correlations would be higher than the correlations with the international item parameter estimates.

The following table presents the correlations within three Chinese-language versions (Hong Kong, Macao, and Chinese Taipei) and between these Chinese versions and the international estimates (OECD, 2009a). The correlations within the Chinese-language versions are substantially higher than their respective correlations with the international item parameter estimates. These correlations reflect a potential language effect.

Table 3: Correlations between Chinese versions and international item parameter estimates

	Hong Kong	Macao	International
Hong Kong			0.82
Macao	0.94		0.85
Chinese Taipei	0.81	0.88	0.75

In addition, it has been found that the correlation between the national item parameter estimates of the two versions in Canada (English and French) is lower than most of the correlations within the English version or within the French version (OECD, 2009a).

These statistics suggest that we cannot dismiss the effect of test translation from our substantive DIF analysis. The items identified as showing DIF through statistical analyses are not necessarily poorly translated items; however, they are good candidates for investigating potential translation problems.

Differential curriculum coverage

PISA aims to provide a curriculum-free measure of students' overall "scientific literacy", which indicates that curriculum should have minimal impact on the functional equivalence of the measure.

To achieve this goal, PISA created a relatively large item pool with nearly 200 items representing a broad spectrum of science content. Those items require students to "identify scientific issues", "explain phenomena scientifically", and "use scientific evidences". Moreover, they also encompassed both "Knowledge Of Science" and "Knowledge About Science". (The former includes understanding fundamental scientific concepts; the latter includes understanding inquiry and the nature of science.)

However, there are reasons to suspect that curriculum may still play a very important role in explaining the detected item DIF. First, OECD reported that students in some countries scored substantially higher in "Knowledge About Science", whereas in other countries students excelled in "Knowledge Of Science". For instance, students in Chinese Taipei and Macao-China scored more than 10 points higher in items measuring "Knowledge Of Science" (OECD, 2007), which suggested that the curriculum has been relatively strong on transmitting specific scientific knowledge in these educational systems.

OECD also reported that students in different countries had substantially different results in the three science competencies ("identifying scientific issues", "explaining phenomena scientifically", and "using scientific evidence"), and in the different content areas such as "Physical systems", "Living systems", and "Earth and space systems" that PISA intended to assess (OECD, 2007). These results might suggest that even though general curricular goals can be similar across systems, specific concepts, skills and behaviors are much more varied.

Furthermore, the most highly emphasized content area represented by the 2006 PISA science item pool is "Nature of Science", accounting for nearly one third of the items. Other content areas with more moderate emphasis include "Measurement in Science" (8 percent), "Human Biology" (9 percent), and "Properties of Matter" (7 percent) (Smithson, 2009). The heavy emphasis on "Nature of Science" is combined with relatively high performance expectations, with items aimed at assessing students' ability to communicate understanding (37 percent), analyze information (22 percent), or make connections or apply to real-world situations (13 percent). In addition, about a quarter (26 percent) of the assessed content measures student recall of science concepts. Interestingly, very few procedural knowledge skills (2 percent) are represented in the PISA science scale (Smithson, 2009). As different countries usually place emphasis on different content domains, skills and procedures, the appropriateness of the content covered and the emphases in the PISA 2006 science item pool remain an open question.

Finally, many of the items in the PISA 2006 Science assessment have a heavy reading load and require a relatively high level of reading comprehension skills (OECD,

2007). It is unclear whether students' science knowledge and skills can be well assessed without any dependence on their literacy skills. PISA presupposes that it is the common goal of the educational efforts in the participating countries to prepare their 15-year-olds to possess the reading comprehension skills necessary to perform well on those items. But it is doubtful whether this presupposition holds in reality.

Based on the evidences and issues discussed above, it seems that the effect of curriculum is too important to be overlooked. Items that display substantial DIF will be scrutinized by content experts in the sampled countries and areas in this study.

Cultural difference

As reviewed in the previous section, differential stimulus familiarity, social desirability, and response style are the three most important aspects to examine when investigating cultural factors as sources of item bias.

Although stimulus familiarity is an important source of method bias in cognitive testing, concerns about differential familiarity in PISA 2006 Science scale may be less because the items were through strict procedures of local paneling and local adaption. However, it is still worthwhile to double check with respect to the particular items that are detected to have substantial DIF.

Social desirability and response style, on the other hand, require special attention because of the heavy emphasis on "attitude toward science" in the PISA 2006 science scale. PISA believes that an important goal of science education is for students to develop interest in and support for scientific inquiry besides acquiring and subsequently applying scientific and technological knowledge for personal, social, and global benefits. That is, a person's scientific literacy includes certain attitudes, beliefs, and motivational orientations that influence their personal actions.

Two formats were employed for the attitudinal items. First, a four-point Likert-type response scale was used. These items did not allow students to opt for a neutral response. A second type, called "match-the-opinion items", asked students to choose from four ordered opinions about an issue, representing different levels of commitment to a sustainable environment.

To ensure the quality of the attitudinal items, the development process also went through carefully designed procedures, as did the cognitive items. The procedures include: the use of skilled professional test development teams from a variety of PISA participating countries; review of the items by experts who have been directly involved in the conceptualization of the underpinning construct definitions; opportunities for the participating countries to review and evaluate the drafted items on multiple occasions; a detailed set of translation and verification protocols that aims at ensuring the conceptual and psychometric equivalence of the items across languages and cultures; and finally,

trial testing activities where students are asked to respond and to reflect upon the meaning of the items.

However, these painstaking procedures cannot eliminate the cross-cultural differences in social desirability and response style entirely. In fact, concerns have been raised because the cognitive scores and attitudinal scores in the PISA 2006 science scale have been found to be negatively correlated at country level (Buckley, 2009). While it is possible that these negative correlations are a result of Simpson's Paradox or aggregation bias, it is highly likely that it is at least partially the result of response style heterogeneity at country level.

An even more alarming indication is that when compared with the scaling results on the pooled international sample, the majority of items were found to have DIF in at least some countries. Table 3.2 shows the number of items that were detected to display DIF in the 51 participating OECD countries compared with the international sample (OECD, 2009a). Hence, it is especially important to take into consideration the effect of social desirability and response style in the analysis on attitude items with substantial DIF items.

Table 4: Summary of DIF analyses on attitudinal items across countries

		In 1-2 Countries	In 3 or More Countries	Total number of Items
Interest Items	# of detected DIF items	22	17	52
Support Items	# of detected DIF items	20	5	37

3.3 Analysis Methods and Procedures

In this study, we focus on the comparisons between (1) U.S. and Canadian students, (2) Hong Kong and mainland Chinese students, and (3) U.S. and Chinese students. These three pairs of comparison groups feature (1) the same languages, similar curricula and cultures, (2) similar languages and cultures, but different curricula, and (3) different languages, curricula, and different cultures. Interestingly, the statistical analyses results can give us some hints about the major causes of DIF between these groups: for instance, if a lot more items display DIF in the Hong Kong-mainland China comparison than in the U.S.-Canada comparison, curriculum difference may be the most prominent cause among these countries. And if the mainland China-U.S. comparison finds more DIF items than the Hong Kong-mainland China comparison, language and culture may be the major contributors for these countries.

However, the identification of the causes of DIF is not likely to be achieved through statistical analysis of the response data alone. Statistical analyses are essential in detecting problematic items, but revealing the causes must rely on substantive investigations as well.

Detailed content analyses were conducted through three parallel investigations, each focusing on the effect of test translation, curriculum and cultural difference, respectively.

Analysis of the Test Translation Effect To find out whether test translation causes DIF in the U.S.-mainland China comparison in this study, items with substantial DIF in both language versions were reviewed by a bilingual person specially trained in translating between the two languages, in terms of the difficulty levels of the key vocabulary as well as the passage as a whole.

It would yield more reliable results if a panel of bilingual experts can be gathered to discuss the problematic items. However, due to the confidentiality requirement regarding the items, the items were only accessible to the author for a very brief period of time, and they cannot be disclosed under any circumstances. Discussing "pseudo-items" (a different item but with similar features) does not make sense for this purpose since it is impossible to carry over all the subtle translation and language features of the original item in a pseudo item. As a result, I can only try my best to review all the problematic items in both language versions, pondering upon the key words, key sentences in the passage, and the general style of the text, and to document all differences that may potentially change the difficulty level of the items. Nevertheless, I consider it to be sufficient for the purpose of this study, as the aim is to identify possible causes of DIF that can be further investigated through more rigorous studies.

Analysis of the Curriculum Coverage Effect The method I applied in this study to examine the curriculum coverage is similar to the OTL survey or the TCMA described earlier in this chapter. Specifically, I consulted a few teachers, science educators and 15-year-old students in each country (consultation questionnaire in Appendix A, B and C) on how well the topics of the DIF items were covered in each country or area.

Content experts from each of the comparison groups, typically experienced science teachers who were very familiar with middle school science curriculum, were asked to rate how well the topics were taught to a typical examinee at the testing age for PISA (OECD, 2009). Their opinions on whether: (1) students have mastered the topics (the topic has been instructed in class, and students have deep understanding of the content, they also have had opportunities to use the knowledge in classroom discussions, solving problems in homework, etc.), (2) have basic understanding about the topic (the topic has been introduced to students in class, but not emphasized. Students may not have had opportunity to learn all detailed aspects of the topic, or they have not had opportunities to use it for problem solving), or (3) have not learned anything about the topic yet, were collected. In addition, a few students were also asked to respond to the same questionnaire so that the confounding factor of teacher perception can be considered.

Besides the topics, the questionnaire also includes one sample item (selected from those officially released by PISA) for the teachers and students to rate the appropriateness of the reading load. Ratings range from "easy to comprehend", "can comprehend with some effort", to "very difficult to comprehend". The purpose of this question is to find out whether differential literacy requirement in these countries and areas may affect the item difficulty levels.

The average ratings from each group were then compared. Discrepancies revealed by the questionnaire are important potential explanations for the detected DIF effects.

Analysis of the Cultural Difference Effect To examine how cultural difference contributes to DIF, I first focused on the social desirability and response style of the attitude items in this study. Formal analysis of response style requires a lot of work and can be another standalone project. As it is not the main angle of this research, a much more simplified procedure was implemented instead.

First, among all the DIF items, I counted the number and percentage of attitude items. Further analysis took place only when attitude items account for a considerably large proportion of DIF items. Secondly, I examined whether one group scored consistently higher or lower (ARS or DARS, and possibly differential social desirability) on those attitude items with substantial DIF. And finally, frequency counts for each scoring category were compared to see whether one group had greater tendency for extremity (ERS) or non-consistency (NCR). Based on these results, patterns were generalized and used to attempt to explain the detected DIF.

Other factors of cultural difference, differential stimulus familiarity and differential social desirability were also explored. We examined the response pattern and discussed the items' contexts with bilingual science educators.

Note that the results from the three analyses, although conducted independently, should not be isolated: DIF in one item might be the synthesized effect of test translation, curriculum, and cultural difference. It is quite possible for DIF to have more than one explanation at a time. Hence, substantive analyses in this study integrated all possible sources to account for item DIF.

References

Allalouf, A. (2000). *Retaining Tranlated Verbal Reasoning Items by Revising DIF Items. .* Paper presented at the AERA, New Orleans, LA.

Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the causes of DIF in translated ver- bal items. *Journal of Educational Measurement, 36*(3), 185–198.

Bachman, J. G., & O'Malley, P. M. (1984). Yea-saying, nay-saying, and going to extremes: Black-White differences in response styles. *Public Opinion Quarterly, 48*(2), 491–509.

Beaton, Martin, M. O., Mullis, I. V. S., Gonzales, E. J., Smith, T. A., & Kelly, D. L. (1996). *Science achievement in the middle school years.* Chestnut Hill, MA: TIMSS International Study Center, Boston College.

Beaton, A. E. (1998). Comparing cross-national student performance on TIMSS using different test items. *International Journal of Educational Research, 29*, 529—542.

Berliner, D. C. (1993). International comparisons of student achievement: A false guide for reform. *National Forum, XXII*(3), 25—29.

Buckley, J. (2009). *Cross-National Response Styles in International Educational Assessments: Evidence from PISA 2006.* Paper presented at the NCES PISA Research Conference.

Chen, C., Lee, S., & Stevenson, H. W. (1995). Response style and cross-cultural comparisons of rating scales among east asian and north american students. *Psychological Science, 6*(3), 170–175.

Cheung, S. K. (1996). Reliability and factor structure of the Chinese version of the Depression Self-Rating Scale. *Educational and Psychological Measurement, 56*, 142-154.

Ercikan, K. (1998). Translation effects in international assessment. *International Journal of Educational Research, 29*, 543–553.

Ercikan, K. & Koh, K. (2005). Examining the Construct Comparability of the English and French Versions of TIMSS. *International Journal of Testing, 5*(1), 23 - 35.

Grisay, A. (2003). Translation procedures in OECD/PISA 2000 international assessment. *Language Testing, 20*(2), 225–240.

Hambleton, R. K., & Kanjee, A. (1995). Increasing the validity of cross-cultural assessments: Use of improved methods for test adaptations. *European Journal of Psychological Assessment, 11*, 147–157.

Hambleton, R. K., & Patsula, L. (1999). Increasing the validity of adapted tests: Myths to be avoided and guidelines for improving test adaptation practices. *Journal of Applied Testing Technology, 1*, 1-12.

Hambleton, R. K., & Patsula, L. (2000). *Adapting Tests for Use in Multiple Languages and Cultures. Laboratory of Psychometric and Evaluative Research Report.*

Marin, G., Gamba, R. J., & Marin, B. V. (1992). Extreme response style and acquiescence among hispanics: The role of acculturation and education. *Journal of Cross- Cultural Psychology, 23*(4), 498–509.

McDonnell, L. M. (1995). Opportunity to learn as a research concept and a policy instrument. *Educational Evaluation and Policy Analysis, 17*(3), 305—322.

OECD (2007). *PISA 2006: Science Competencies for Tomorrow's World* (Vol. 1).

OECD (2009). *PISA 2006 Techinical Report.*

Price, L. R., & Oshima, T. C. (1998). *Differential Item Functioning and Language Translation: A Cross-National Study with a Test Developed for Certification.* Paper presented at the AERA, San Diego, CA.

Schmidt, W. H., Jakwerth, P. M., & McKnight, C. C. (1998). Curriculum sensitive assessment: Content does make a difference. *International Journal of Educational Research, 29*, 503-527.

Siegler, R. S. (1998). *Children's Thinking*. Upper Saddle River, NJ: Prentice Hall.

Sireci, S. G., & Swaminathan, H. (1996). *Evaluating Translation Equivalence: So What's the Big DIF?* Paper presented at the AERA, Ellenville, NY.

Smithson, J. (2009). *Descibing the Academic Content of PISA Mathematics and Science Item Pools.* Paper presented at the the NCES PISA Research Conference.

Vijver, F. J. R. v. d., & Poortinga, Y. H. (1997). Towards an Integrated Analysis of Bias in Cross-Cultural Assessment. *European Journal of Psychological Assessment, 13*(1), 29–37.

Watkins, D., & Cheung, S. (1995). Culture, gender, and response bias: An analysis of responses to the Self-Description questionnaire. *Journal of Cross-Cultural Psychology, 26*(5), 490.

Westbury, I. (1993). American and Japanese achievement...again. *Educational Researcher, 22*(3), 21—25.

Wolf, R. M. (1998). Validity issues in international assessments. *International Journal of Educational Research, 29*, 491—501.

Wu, A. D., & Ercikan, K. (2006). Using Multiple-Variable Matching to Identify Cultural Sources of Differential Item Functioning. *International Journal of Testing, 6*(3), 287 — 300.

Chapter 4: Results

4.1 General Psychometric Properties of the Instrument

Before we delve into DIF analysis, it is informative and necessary to review the general psychometric properties of the instrument. Specifically, we examined the reliabilities, person and item distributions, and item fit statistics for each group from separate calibration results. These statistics not only reveal how well the assessment functions within each group but also help us assess whether the scaling model holds across all the groups in this study.

Reliabilities Reliability is a fundamental requirement for any measures to be used for any kinds of inferences or decisions. When calibrated separately, the reliabilities for the Science assessment for Canadian, U.S., Hong Kong, and mainland Chinese students are 0.91, 0.93, 0.92, and 0.89, respectively. The reliabilities are quite high and warrant further uses of the test results, although with the large number of items one might expect a higher reliability.

Item and Person Distributions By examining item and person distributions, we can see whether the item difficulty levels match the examinee ability levels to provide optimum information. For each of the four separate calibrations, the mean item difficulty was set at a common point (i.e., zero). Table 5 shows the item distribution statistics of the four groups. And Table 6 summarizes the person distribution statistics.

Table 5: Item Distribution Statistics

	For Canadian Students	For U.S. Students	For HK Students	For Mainland Chinese students
Easiest Item	-2.25	-2.11	-2.16	-2.99
Most Difficult Item	1.91	2.19	2.54	2.90

Table 6: Person Distribution Statistics

	N (number of students)	Mean	Standard Deviation	Minimum	Maximum
Canadian Students	17555	0.40	0.67	-1.84	2.80
U.S. Students	5611	0.25	0.62	-1.79	2.42
HK students	4646	0.77	0.68	-1.50	3.12
Mainland Chinese	4892	0.99	0.56	-0.87	2.83

We found that, for Canadian students, the item difficulties were between -2.25 and 1.91 logits. The person distribution had a mean of 0.40 and a standard deviation of 0.67. The Wright map[1] (Wilson, 2005), shown in Figure 1, suggests that the person

[1] The first column on the left-hand side marks the logits. The “X”s on the left side of the vertical line are the locations of respondents on the proficiency scale. In this graph, each 'X' represents 114 examinees. It is

distribution was approximately normal and the items covered most of the ability spectrum, with just a few items that were a bit too easy for the students.

For U.S. students, the item difficulties spanned from -2.11 to 2.19. The person distribution had a mean of 0.25 and a standard deviation of 0.62. The Wright map (Figure 2) shows that the person distribution was also normal, and the whole range was well covered by the items.

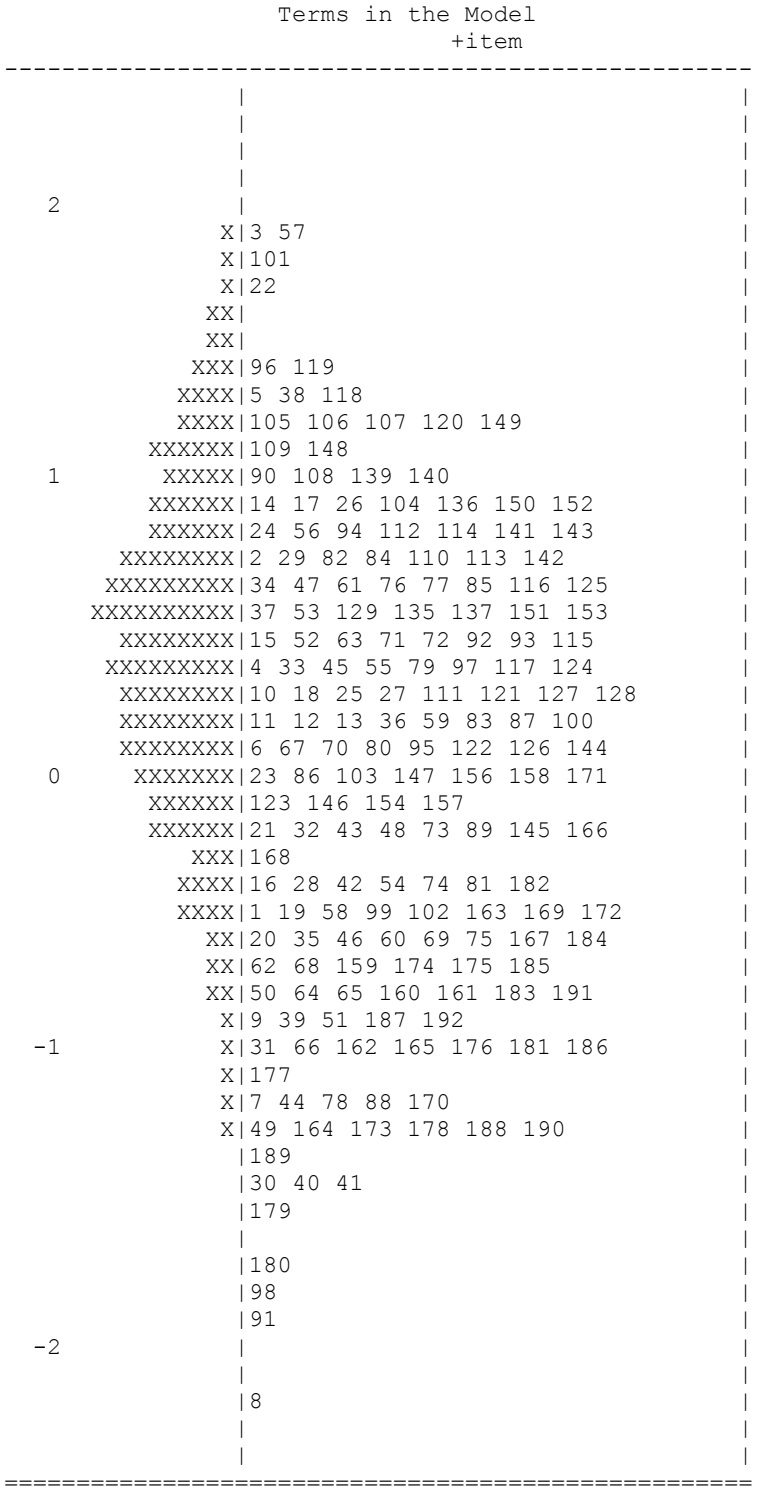
For Hong Kong Students, the item difficulty ranged from -2.16 to 2.54, whereas the mean person ability was 0.77 with a standard deviation of 0.68. Figure 3 suggests that the person distribution was quite normal. The items covered the ability spectrum well, although there were some items that appeared to be too easy for the students.

For mainland Chinese students, the easiest item had a difficulty of -2.99 logits, while the hardest item had a difficulty of 2.90 logits. The mean person ability was 0.99 with standard deviation 0.56. The Wright map (Figure 4) shows a normal person distribution and a good coverage of the ability span. However, there were more items that appeared to be too easy for the students, which might explain the relatively lower reliability for Chinese students.

---

in the shape of an on-the-side histogram. On the right-hand-side of the figure, under "+item", are the locations of items, denoted by the item numbers.

Figure 1: Wright Map for Canadian Student



- Each 'X' represents 114 cases

Figure 2: Wright Map for U.S. Students

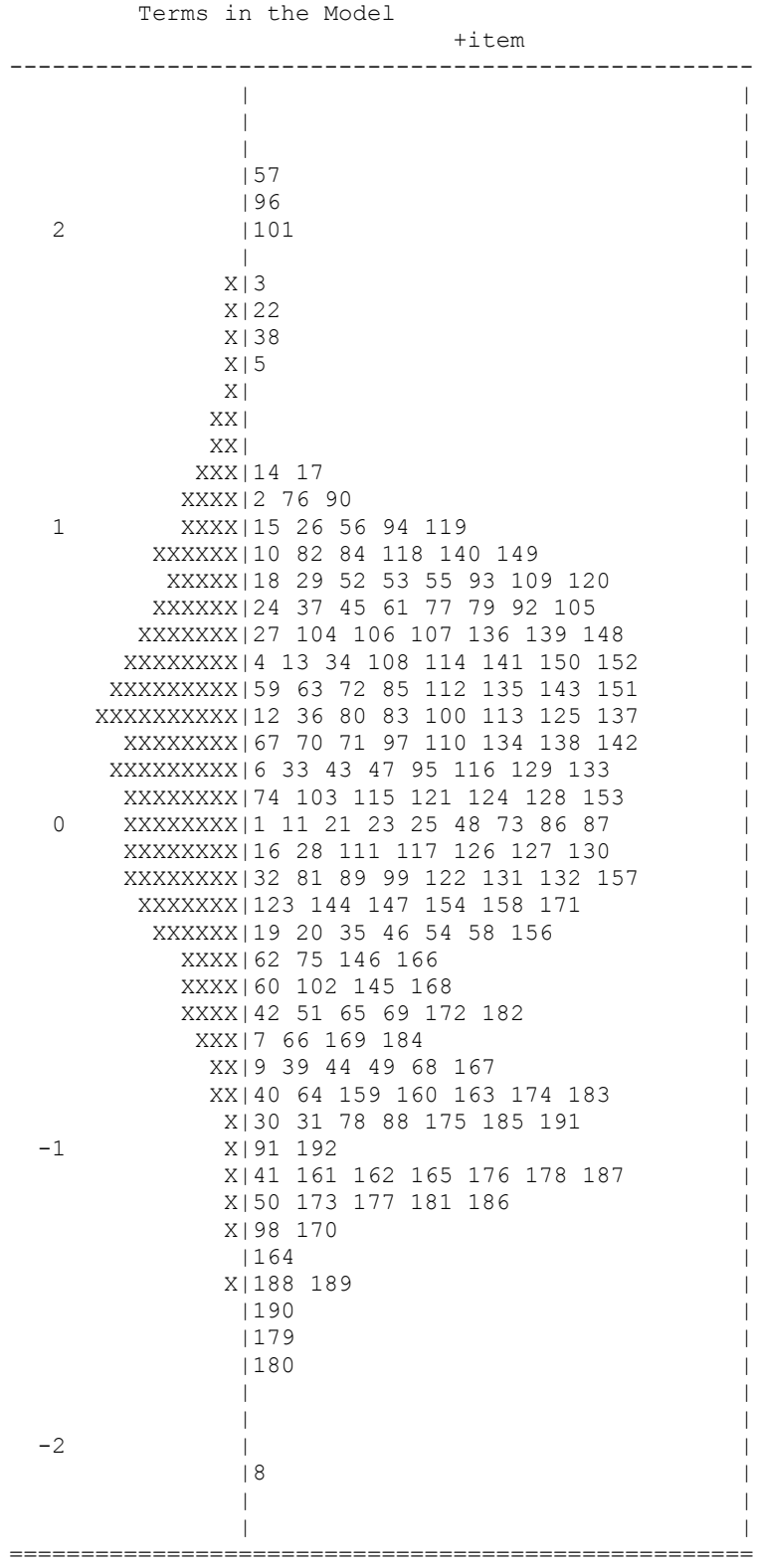


Figure 3: Wright Map for HK students

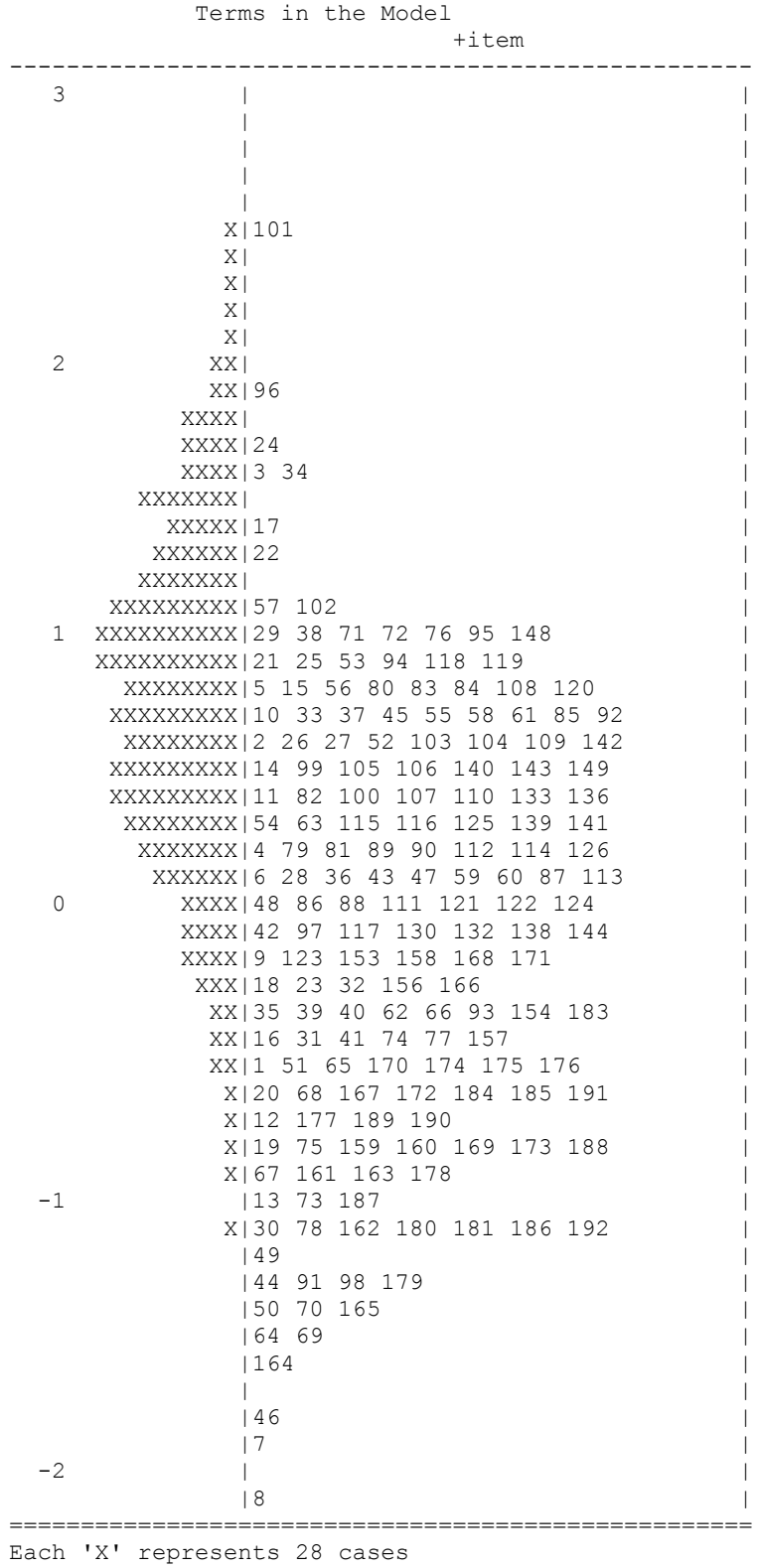
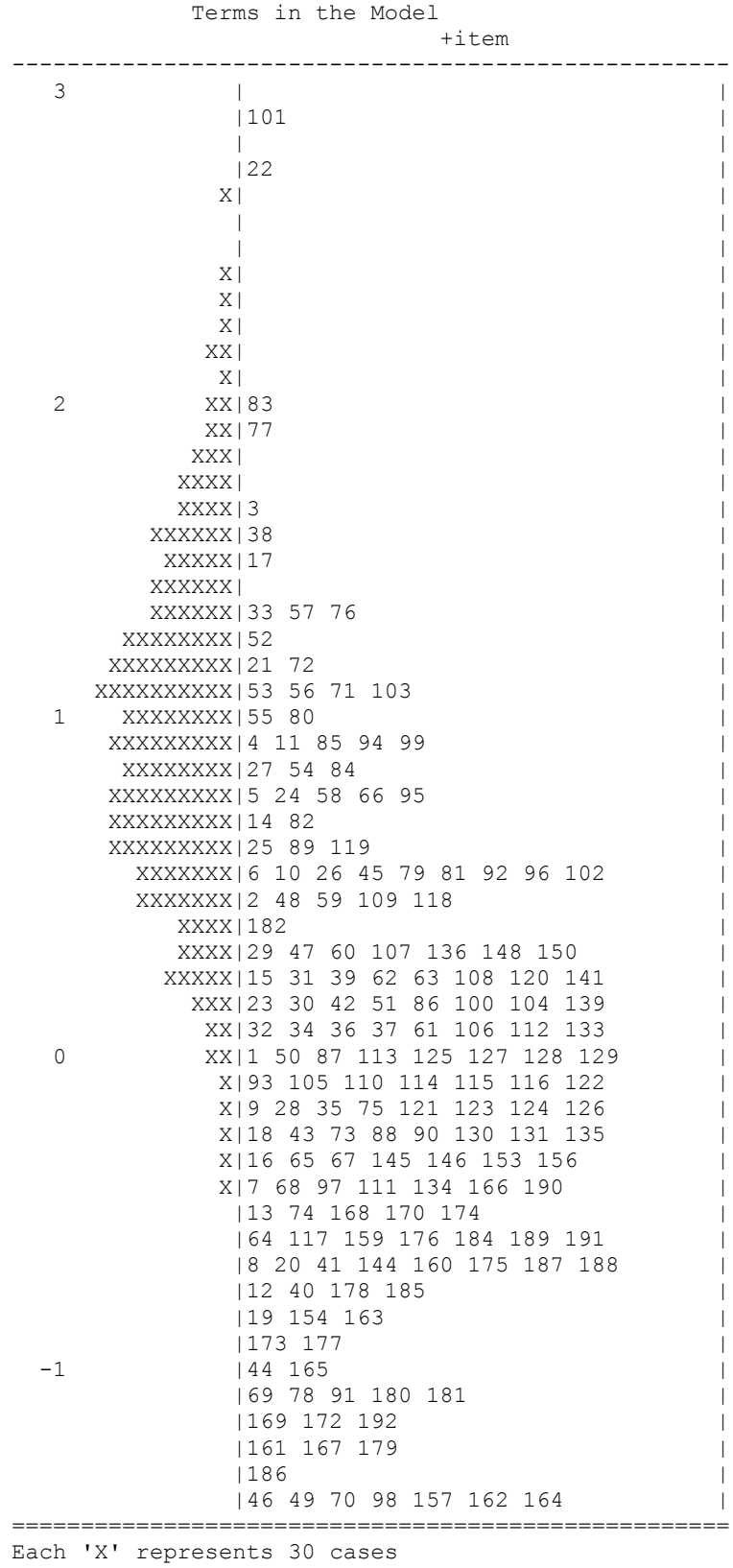


Figure 4:Wright Map for Mainland Chinese Students



Item Fit Statistics The item fit statistics is an important index of how well the response data meets the expectation of the measurement model, but good model-data fit does not mean the item is DIF free. It is important to make a distinction between DIF and misfit. An item can fit the model's expectation reasonably well within group, but can have different difficulty estimates for different groups (which is DIF).

Item misfit can occur due to many causes, such as local dependence, multidimensionality, inappropriate discrimination power, guessing, etc (Wang, 2008). If an item is found to be misfitting, it should be removed from subsequent DIF analysis. This procedure ensures that item parameter estimates obtained using the Rasch family scaling models are meaningful and can be used as evidences of further analyses or discussions.

One of the most commonly used item fit indices is the weighted mean square statistics (sometimes called the infit statisitics). It has been suggested that values between 0.75 (=3/4) and 1.33 (=4/3) indicate reasonably good item-model fit (Wilson, 2005).

Our result shows that the items met the expectation of the Rasch model very well for all four groups. The infit statistics of the items were within the range of 0.75 to 1.33. This is not surprising as PISA items are scrutinized for misfit. Hence the following DIF analyses included all the items in the PISA 2006 Science assessment.

4.2 U.S. vs. Canadian Students

Unidimensional DIF analysis results

We first examined the functional equivalence of the test between U.S. and Canadian students under the unidimensional framework. The U.S. and Canadian students' response data were scaled together using the RCML DIF-detection model. We found that the reliability was very high (0.92). The item difficulties spread from -1.81 to 2.33 logits, covering the ability range quite well (shown in Figure 5[2]). The infit statistics of all the items ranged between 0.83 and 1.25, suggesting that all items fit the model's expectation quite well.

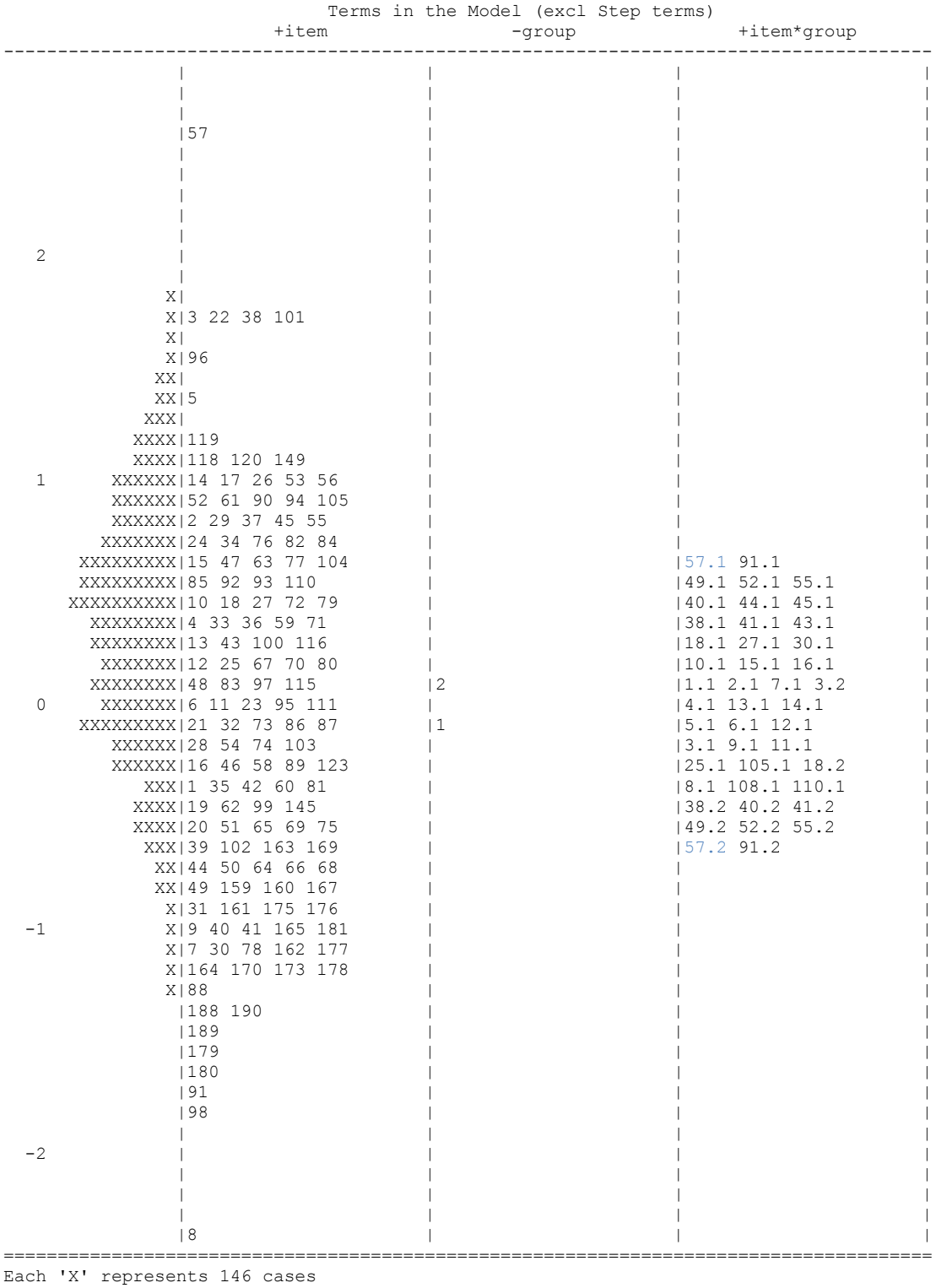
[2] As in Figure 1, the first column from the left-hand side marks the logits. The locations of examinees are denoted by "X"s, each representing 146 cases in this figure. The item difficulties are shown in the second column, under "+item". The third column, under "-group", shows the mean location of the two groups of children, "1" denoting U.S. students and "2" denoting Canadian students. The column on right-hand side, under "+item*group", presents the DIF parameters for the two groups. The first number is the item number. The suffix refers to the group identity. '1" refers to the U.S. group, and "2", the Canadian group. For example, "57.1" refers to the DIF parameter of item 37 for U.S. students. For each item, the DIF parameters for the two groups are symmetric to 0. The DIF effect size of an item is the difference between the parameters for the two groups. For instance, the DIF effect size for item 57 is the difference between "57.1" and "57.2". Note that not all items are shown in this figure due to limited space.

The analysis found 41 items (21.3%) with substantial DIF (class B and C DIF). The Chi-square statistic test showed that the item-by-group interactions were statistically significant for all 41 items. The largest item-by-group interaction was as high as 1.40. Table 7 presented the number of items identified as manifesting DIF in favor of Canadian and U.S. students. Specifically, more items, 28 out of the 41, favored Canadian students, and half of them (14 items) had effect sizes larger than 0.638 (Class C DIF). The remaining 13 items favored U.S. students, with only 1 Class C DIF item.

Table 7: Unidimensional DIF Analysis between Canadian and U.S. Students

	Number of Class B DIF Items	Number of Class C DIF Items	Total Number of Items
Favoring Canadian Students	14	14	28
Favoring U.S. Students	12	1	13
Total Number of Items	26	15	41

Figure 5: Wright Map for Canadian and U.S. students



Multidimensional analysis

Since the percentage of DIF items exceeded 10%, a multidimensional procedure was followed. As elaborated in Chapter 2, a three-dimensional model was adopted, with 103 items loading on the "cognitive" dimension, 52 items on the "interest" dimension, and 37 items on the "support" dimension. The correlations among the three dimensions for Canadian and U.S. students displayed a similar pattern found from the international calibration (presented in Chapter 2). Table 8 showed that the correlation between the "cognitive" and "support" dimension was quite high (r=0.82, higher than international calibration results of 0.60), whereas the correlation between the "cognitive" and "interest" dimension" was almost zero (r=0.08, very close to the international average of 0.06). The two attitude dimensions ("interest" and "support") had a moderately high correlation (r=0.52).

Table 8: Correlations among the Three Dimensions for Canadian and U.S. Students

	Cognitive Dimension	Interest Dimension
Interest Dimension	0.08	
Support Dimension	0.82	0.52

The reliabilities for dimensions "cognitive", "interest" and "support" were 0.84, 0.73 and 0.79, respectively. (It is not surprising that the reliabilities of the two attitude dimensions are lower since there are fewer items in those dimensions.)

7 items were found to have misfit. But in general, the multidimensional model fit the data significantly better than the unidimensional one, with a total reduction of 1761 in the model deviance statistics. The Chi-square test was significant with 3 degrees of freedom.

Using the same criteria of 0.426 for substantial DIF, the multidimensional DIF-detection model (MRCMLM) found 10 items (5.2%) with substantial DIF. Among them, only 2 had effect sizes greater than 0.638 (Class C DIF). Table 8 summarized the number of DIF items favoring each group. We can see that more items (7 out of the 10 items) favored U.S. students under the multidimensional model.

Table 9: Multidimensional DIF Analysis between Canadian and U.S. Students

	Number of Class B DIF Items	Number of Class C DIF Items	Total Number of DIF Items
Favoring Canadian Students	2	1	3
Favoring U.S. Students	6	1	7
Total Number of Items	8	2	10

Comparing the results of the unidimensional and multidimensional approaches, it is quite obvious that the number of DIF items reduced drastically (from 41 to 10 items) under the multidimensional framework. Moreover, the magnitude of DIF also reduced significantly from 1.40 as the maximum in the unidimensional analysis to 0.65 as the maximum in the multidimensional analysis.

One important observation is that the earlier unidimensional analysis found altogether 12 DIF items in the two attitude dimensions (10 "interest" items and 2 "support" items, and that all 12 items favored U.S. students). However, under the multidimensional framework, none of these 12 items displayed DIF any longer. That is, when calibrated using the multidimensional model, all the detected DIF resided in the "cognitive" dimension.

These results suggested that multidimensionality explained a fairly large amount of DIF between the Canadian and U.S. students. The three subscales were so distinct that the unidimensional assumption did not hold. And falsely assuming unidimensionality caused DIF.

Systematic DIF

Although the number of DIF items was smaller than 10%, we still compared person ability estimates obtained from a model in which DIF items were excluded with those obtained from the original model in which DIF items were not excluded, in order to make sure that the PISA 2006 Science assessment does not contain systematic DIF for U.S. and Canadian students.

Our results showed that the correlation between the two person measures was extremely high ($r=0.99$). Although the paired-sample t-test was statistically significant ($p<0.01$), the effect size was trivial ($d=0.006$). Previous research has suggested that t-test can be significant even with a negligible difference when the sample size is large. But since the effect size was so small, it seems safe to assume that there was no systematic DIF between the two groups of students.

This finding is hardly surprising since U.S. and Canadian students are relatively similar: their test languages are both English, their educational systems are more similar comparing to those in other continents, and they also have relatively similar cultures.

To sum it up, when comparing the PISA 2006 Science assessment results between Canadian and U.S. students, making further inferences or decisions, the response data needs to be calibrated using a multidimensional model, or each of the three subscales should be compared independently.

As DIF is not a serious threat to the cross-country validity for these two groups of students under a multidimensional framework, detailed content analysis is not conducted in this case.

4.3 Mainland Chinese vs. Hong Kong Students

Unidimensional DIF analysis results

To investigate whether DIF is a problem between mainland Chinese and Hong Kong students, response data from these two groups of students were first scaled together using the unidimensional RCML model. The reliability was 0.89. All items seemed to fit the expectation of the model fairly well, with infit statistics within the critical values of 0.75 and 1.33. Item difficulties ranged from -1.62 to 2.73. The Wright map (Figure 6[3]) showed that quite a few items appeared to be too easy for the students, which might be the reason why the reliability was lower than that of the U.S. and Canadian sample.

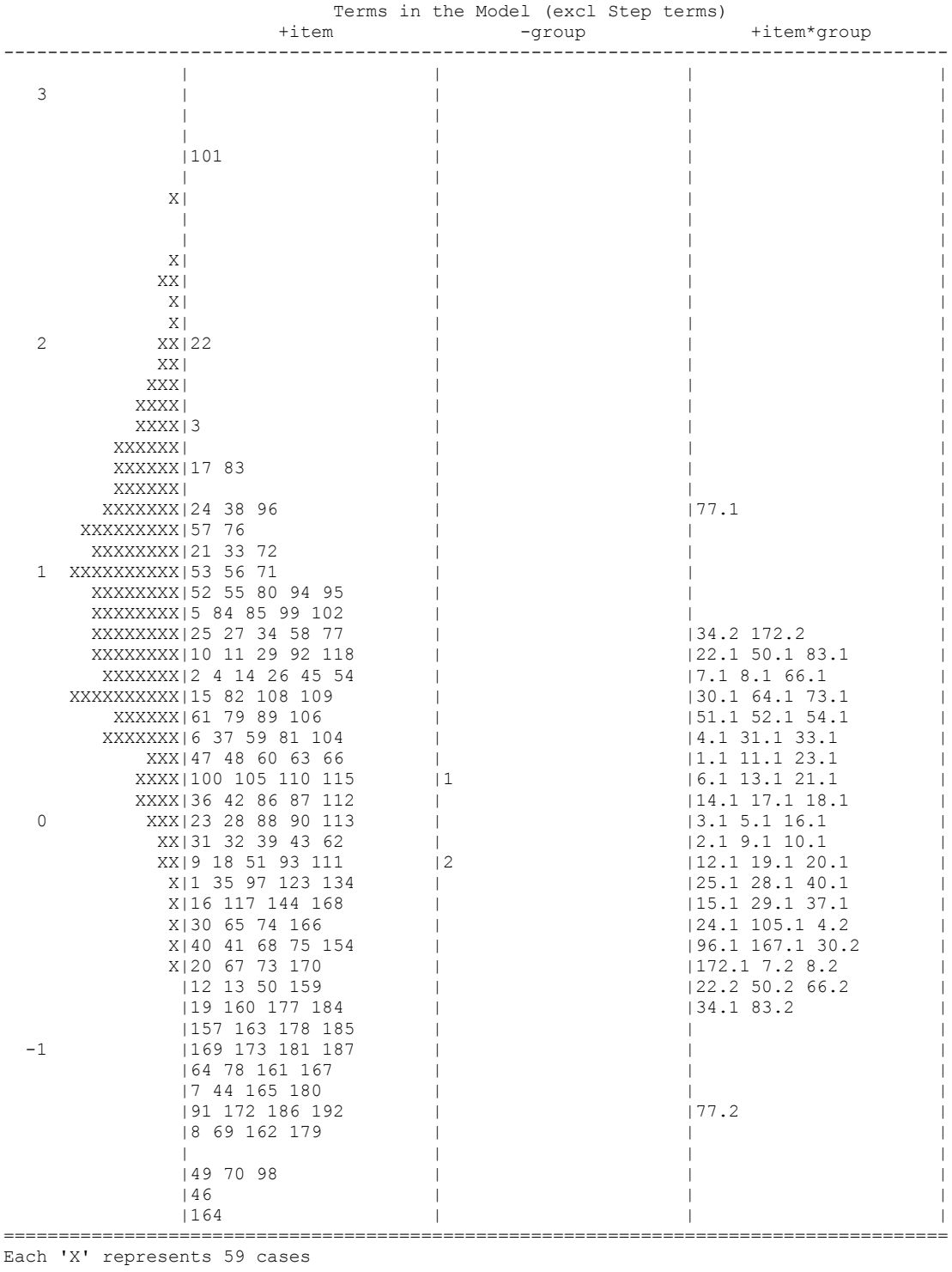
The analysis revealed that there were 76 items (an alarming 39.6%) with substantial DIF. Among them, 39 items favored mainland Chinese students and 37 favored Hong Kong students. About half of the DIF items (33 items) had effect sizes larger than 0.638 (Class C DIF), with 13 favoring mainland Chinese students and 20 favoring Hong Kong students. Table 10 summarized the number of DIF items favoring each group. In addition, the largest group difference was 2.52 logits, a very large value.

Table 10: Unidimensional DIF Analysis between Mainland Chinese and HK Students

	Number of Class B DIF Items	Number of Class C DIF Items	Total Number of DIF Items
Favoring Mainland Chinese	26	13	39
Favoring HK Students	17	20	37
Total Number of Items	43	33	76

[3] This figure is similar to Figure 5. Here, each "x" represents 59 cases. In addition, under "-group", "1" denotes mainland Chinese students, and "2" denotes Hong Kong students. Accordingly, under "+item*group", the suffix "1" denotes DIF parameter for mainland Chinese students, and "2" for Hong Kong students. All other symbols hold the same interpretations as in Figure 5.

Figure 6: Wright Map for Mainland Chinese and U.S. students



Multidimensional DIF analysis results

Because the unidimensional analysis detected quite a high percentage of DIF items, a multidimensional analysis was followed. The same three-dimensional model as in the U.S.-Canadian analysis was adopted.

Table 11 summarized the correlations between the three dimensions for mainland Chinese and Hong Kong students. Dimension "cognitive" and dimension "support" had a relatively high correlation of 0.64, while the correlation between dimension "cognitive" and dimension "interest" was again extremely low (r=0.03). (The general pattern is also in accordance with that obtained from the international calibration.)

Table 11: Correlations among the Three Dimensions for Mainland Chinese and HK Students

	Cognitive Dimension	Interest Dimension
Interest Dimension	0.03	
Support Dimension	0.64	0.49

The reliabilities of dimension "cognitive", "interest" and "support" were 0.80, 0.47, and 0.64, respectively. The reliabilities of the two attitude dimensions were noticeably lower than that of the cognitive dimension. Fewer items in those two dimensions might be part of the reason. However, the reliability of dimension "interest" was even lower than that of dimension "support", although there were more items in the "interest" dimension. In addition, it was also much lower than those of the U.S. Canadian sample. This suggests that the items measuring student "interest in Science" may not perform as well for these two groups of students as for their peers in North America.

6 items were detected to have model-data misfit. It was a very small proportion (3.1%) and should not raise a big concern. In addition, the overall model-data-fit improved significantly comparing with the unidimensional one: the deviance statistic reduced by 9262.92. The Chi-square test was significant with 3 degrees of freedom.

The multidimensional analysis found that the number of items with substantial DIF reduced significantly from 76 (39.6% in the unidimensional analysis) to 36 (18.8% in the multidimensional analysis), which was almost half as many as in the unidimensional analysis. The magnitude of DIF also reduced from a maximum of 2.52 to a maximum of 2.24 logits, not so great a reduction as for the previous analysis.

Table 12 summarized the numbers of Class B and Class C DIF items favoring each group. Specifically, just about half of the DIF items favored each group. And among the 20 Class C DIF items, more (12 items) favored mainland Chinese students.

Table 12: Multidimensional DIF Analysis between Mainland Chinese and HK Students

	Number of Class B DIF Items	Number of Class C DIF Items	Total Number of DIF Items
Favoring Mainland Chinese	7	12	19
Favoring HK Students	9	8	17
Total Number of Items	16	20	36

Further more, as observed in the previous U.S.-Canadian analysis, multidimensional analysis again eliminated all DIF in the two attitude dimensions. 30 items, 21 in dimension "interest" and 9 in dimension "support", that displayed substantial DIF (all favoring mainland Chinese students) in the original unidimensional analysis no longer had DIF under the multidimensional framework. These items constituted 75% of the total reduction in the number of DIF items. This suggested that multidimensionality was an importance cause of the DIF found in the unidimensional analysis.

Systematic DIF

Although the number of DIF items reduced dramatically under the multidimensional framework, it still exceeds 10%. Moreover, among the remaining 36 DIF items, 20 (55.6%) have large DIF effects (Class C DIF). We would naturally suspect that DIF might exist at test level.

We first compared the overall person ability estimates obtained from a model in which DIF items were excluded with those obtained from the original model in which DIF items were not excluded. The correlation was very high (r=0.98). The paired t-test was statistically significant, although the actually difference was quite small (d=0.06).

In addition, we compared the two person ability estimates of the "cognitive" dimension alone. The correlation was a bit lower than that of the overall ability estimates, but was still fairly high (r=0.96). The t-test was again significant with a very small difference of 0.03.

As these t-tests may be significant due to the large sample sizes but not the actually differences, it seems that DIF did not have a serious effect on ranking ordering mainland Chinese and Hong Kong students.

However, as the number of DIF items, especially the number of Class C DIF items was quite large, the interpretations of the test scores may be qualitatively different for these two participating groups. Specific suggestions or inferences that are based on the use of specific items scores will not be appropriate.

In summary, we strongly suggest that a multidimensional approach be adopted when analyzing data for these two groups of students. In addition, although the test results may be safely used to make general overall comparisons under a multidimensional

framework, functional equivalence does not hold at item level with about one fifth of the items exhibiting substantial DIF.

Detailed content analysis was followed to explore more potential causes of the remaining DIF so that future test development efforts can take those factors into consideration, and in turn produce test scores that can provide specific information at item level.

Causes of DIF between mainland Chinese and HK Students

As stated in Chapter 3, investigations on the causes of DIF focus on three main aspects: (1) the effect of test translation, (2) differential curriculum coverage, and (3) cultural difference.

Test Translation Test translation is not a big concern for Chinese and Hong Kong students since they use very similar languages. First, in terms of their spoken languages, the two groups of students speak two dialects in the Chinese language family. The examinees from mainland China speak standard Mandarin, while most Hong Kong students speak Cantonese. However, standard Mandarin has become one of the official languages in Hong Kong since 1997. In fact, it is now one of the main languages of government, the media and education now. As a result, Hong Kong students can usually understand and speak Mandarin quite well. Furthermore, Cantonese and Mandarin are two closely related varieties of the Chinese language. Although they might be mutually intelligible phonetically, the grammars of the varieties share most of the major traits (DeFrancis, 1986).

As far as the written languages were concerned, Hong Kong students took the test in Traditional Chinese, while mainland Chinese students took it in Simplified Chinese. Traditional Chinese and Simplified Chinese are two different writing systems for the Chinese characters. The traditional system, used in Hong Kong, Taiwan, Macau and most Chinese speaking communities outside mainland China, takes its form from standardized character forms dating back to the late Han dynasty. The Simplified Chinese character system, developed by the People's Republic of China in 1954 to promote mass literacy, simplifies most complex traditional glyphs to fewer strokes, many to common “caoshu” shorthand variants. ("Chinese Language," 2010) Although the simplification aimed to alleviate the burden of learning characters, research shows that time needed to master the characters is not significantly reduced (DeFrancis, 1986). Grammatically, the two systems are identical.

Hence, the two versions of the Science assessment, the Simplified Chinese and the Tradition Chinese versions, differ only in the writing of the Chinese characters. The impact of language is minimal, and test translation should not be an issue here.

Curriculum Coverage As the educational system in Hong Kong is closely modeled on the one used in UK ("The Hong Kong education and schooling system explained," 2005), it is not surprising that its curricula differ from those used in mainland China. To find out whether differential curriculum coverage is a major cause of DIF in the PISA Science assessment, I enlisted the help of two expert teachers and two "average" fifteen-year-old students from each area, and consulted them on how well the topics of those DIF items were covered in their curricula.

As explained in Chapter 3, the teachers and students were asked to rate whether students have mastered the topics, have basic understanding about the topics, or have not learned anything about the topics yet. They were allowed to discuss the questions with their colleagues or classmates. (The consultation question list for Hong Kong teachers and students is shown in Appendix A, and the one for mainland Chinese teachers and students is shown in Appendix B. Note that there are more questions for mainland Chinese teachers and students because the list also includes topics in DIF items found between mainland Chinese and U.S. students.)

The ratings from each group were then compared. The results showed that teacher ratings and student ratings were consistent in most cases. But there were a few discrepancies on how well students have learned the topics. When teacher perceptions and student perceptions don't agree, average ratings were considered.

We found that curriculum difference could explain some of the detect DIF. Specifically, among the 19 DIF items that favored mainland Chinese students, the contents of 5 items were better covered in the curricula in mainland China than in those of Hong Kong students. And 2 out of the 17 DIF items favoring Hong Kong students contained topics that were better covered in Hong Kong's curricula. For instance, it was reported that students in mainland China should have mastered the "law of buoyancy" by the time they took the assessment, while in Hong Kong the topic would not have been introduced yet. Another example is that Hong Kong students were expected to know "the function of human heart" very well, while in mainland China the topic had not been discussed in details yet.

Besides the DIF items that can be directly accounted for by differential curriculum coverage or emphases, we noticed that most DIF items measuring "Knowledge About Science" (defined as understanding of inquiry and the nature of science) favored Hong Kong students. Items on "how to design scientific experiments", "how to collect scientific evidences" and "how to explain scientific phenomena" were more often than not easier for Hong Kong students than for mainland Chinese students. On the other hand, most of the DIF items favoring mainland Chinese students measure "Knowledge Of Science" (defined as understanding of fundamental scientific concepts and facts). This suggested that the curricula in mainland China might be relatively strong on transmitting specific science knowledge whereas the curricula in Hong Kong might have stronger emphases on scientific inquiry. This trend was not shown from the consultation results. One possible explanation is that although the terms "master level" and "basic level" were defined and explained to the teachers and students I consulted,

their perspectives on how well the topics were taught/learnt in class still varied to some extent. It is also possible that the differences could not be well reflected because their responses are forced between two levels of mastery.

In addition to the item contents, I also collected teacher and student opinions on the appropriateness of the reading load of a sample item (a released PISA item) to see whether the literacy requirement affects the item difficulty levels for the two groups of students. Interestingly, teachers from both groups thought that the item's reading load was moderate and their students should be able to comprehend most of it with some effort. But student opinions differed. While students in Hong Kong felt that they had no problem understanding the text at all, students in mainland China thought that comprehending the text required some effort.

So I examined the length of the text and the complexity of the graphs of the DIF items and found that they vary quite a lot. There were no clear evidences that items with longer texts or more figures were more likely to favor Hong Kong students. So it seems that the teachers' perceptions were reliable here and there was no differential literacy requirements. The reason why student opinions differed could be that the students who responded to the consultation have different literacy abilities. Surveying a larger pool of teachers or students may give us deeper insights on this issue.

Nevertheless, comparing the curriculum sources of the item contents showed a differentiated match between item and curriculum for 7 items (19.4% of all the detected DIF). Furthermore, we suspected that "Know About Science" was better conveyed in Hong Kong's curricula, while curricula in mainland China covered more "Knowledge of Science" topics assessed in PISA 2006 Science test. Differential curriculum coverage does seem to be an important cause of the detected DIF between mainland Chinese and Hong Kong students.

Cultural Differences To examine whether differential social desirability and response style contribute to DIF found between the two groups of students, we first looked at DIF items in the two attitude dimensions. As noted earlier, under the multidimensional framework, none of the attitude items exhibited substantial DIF, which suggested that social desirability and response style were not causes of DIF here.

However, as noted earlier, the 30 attitude items (21 "interest" items and 9 "support" items) that displayed DIF in the unidimensional analysis all favored mainland Chinese students. In fact, the average item difficulties of the two attitude dimensions were lower for mainland students. The difference was especially large for the "interest" dimension: when calibrated separately, the mean difficulty of this dimension was -0.04 for mainland students but 0.32 for Hong Kong students.

This pattern might suggest an acquiescence response style (ARS, defined as a tendency to agree with items regardless of actual attitude) for mainland Chinese students or a disacquiescence response style (DARS, defined as a tendency to agree with items regardless of actual attitude) for Hong Kong students (Buckley, 2009). The difference is

also possibly the result of differential social desirability, that is, it might be more socially desirable to show strong interest in Science in mainland China than in Hong Kong. Finally, the pattern could reflect actual difference in "interest" and "support" between the two groups of students, which is know as differential impact, but not DIF (Wilson, 2005). The current study does not collect evidence for what the true reasons are for the different difficulty levels of the two subscales. Future research may investigate this issue through surveys on student perceptions.

Secondly, we examined whether the DIF items in the "cognitive" dimension contained contents that was more familiar to one of the groups. Only one DIF item, which talked about a magnetic hover train, was suspected to be more familiar to mainland Chinese students since there were one in use in Shanghai and another one under construction in Zhejiang Province at the time of the test. And indeed this item was found to favor mainland Chinese students. The remaining DIF items could not be explained by any differential content familiarity we could detect.

In a word, cultural difference does not seem to be a major cause of DIF between mainland Chinese and Hong Kong students. This is only natural since the two groups of examinees have very similar cultures and social practices.

In summary, not all the DIF items can be explained by the three potential causes discussed above. However, it seems that differential curriculum coverage is the most important reason we can find so far for these two groups of students.

4.4 U.S. vs. Chinese Mainland

Unidimensional DIF analysis results

U.S. and Chinese students differ drastically in terms of language, curriculum and culture. Our hypothesis is that DIF will be most serious for these two groups of students. To investigate this issue, we first calibrated the data using the unidimensional RCML DIF-detection model. The reliability turned out to be 0.89. It is the lowest in the three analyses in this study. All the items had good model-data fit. The item difficulties spread from -1.81 to 2.33 logits. The Wright map[4] in Figure 7 showed that the items covered the ability spectrum quite well although there were a few items on the lower end of the continuum that appear to be too easy for the students.

The analysis identified 95 items as showing substantial DIF, which was almost half of all the items (49.5%) in the test. Moreover, among them, 62 (65.3%) had effect

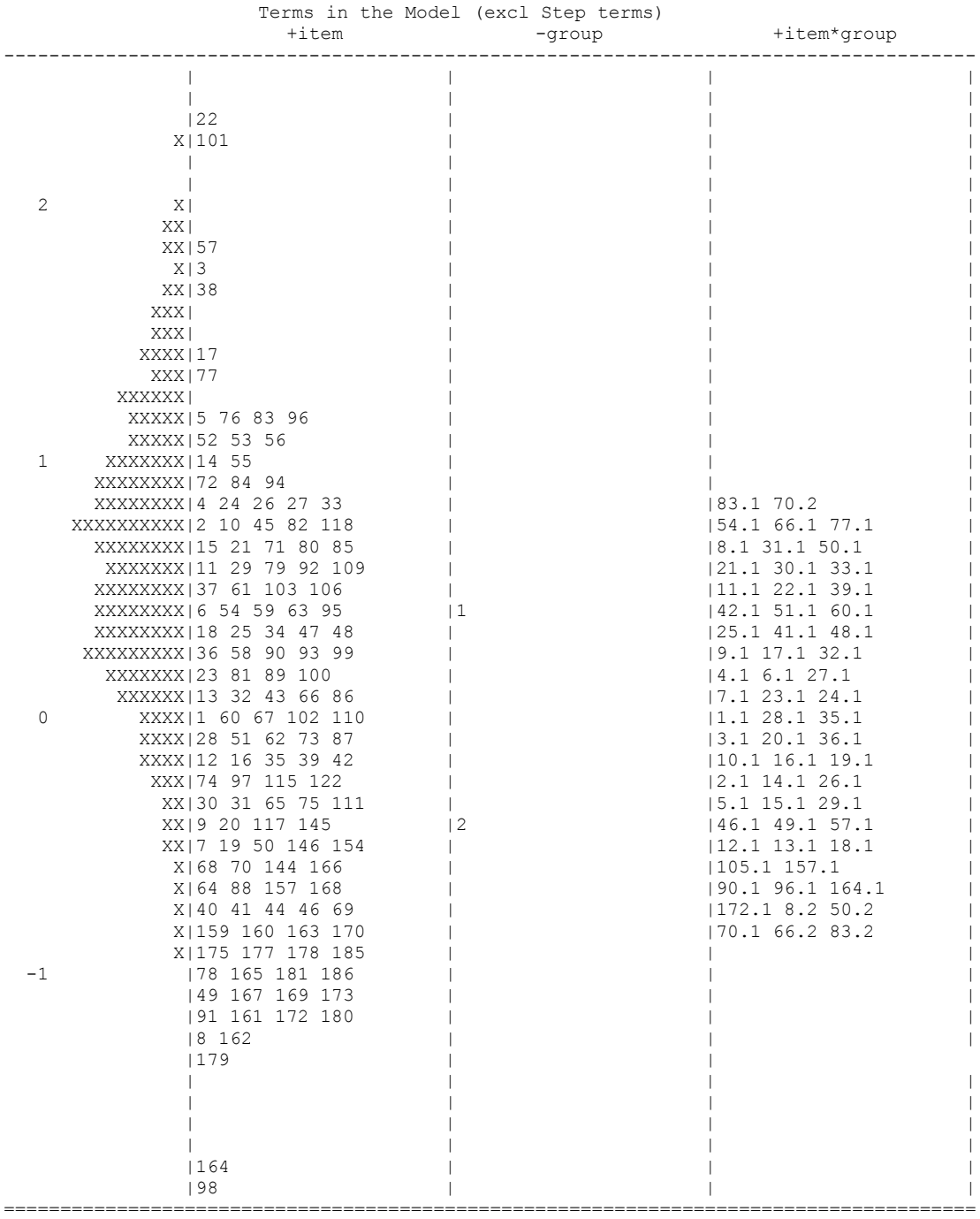
[4] In this figure, each "x" represents 67 cases. In addition, under "-group", "1" denotes mainland Chinese students, and "2" denotes U.S. students. Accordingly, under "+item*group", the suffix "1" denotes DIF parameter for mainland Chinese students, and "2" for U.S. students. All other symbols hold the same interpretations as in Figure 5 and Figure 6.

sizes larger than 0.638. The largest between-group difference was as high as 1.72 logits. Table 13 summarized the number of DIF items in favor of each group. Specifically, about half of the DIF items favored Chinese students. Among the 62 Class C DIF items, 8 more items favored U.S. students.

Table 13: Unidimensional DIF Analysis between Chinese and U.S. Students

	Number of Class B DIF Items	Number of Class C DIF Items	Total Number of DIF Items
Favoring Chinese Students	22	27	49
Favoring U.S. Students	11	35	46
Total Number of Items	33	62	95

Figure 7: Wright Map for U.S. and Chinese Students



Multidimensional DIF analysis results

Compared to 39.6% between Hong Kong and mainland Chinese students, and 21.3% between Canadian and U.S. students, the 49.5% of DIF items between Chinese and U.S. students found in the unidimensional analysis was astonishingly high. Based on the results of the two previous multidimensional analyses, we have reason to suspect that multidimensionality could be an important cause of DIF detected from the unidimensional analysis.

The same three-dimensional MRCML DIF-detection model was applied again. The correlations between the three dimensions were shown below in Table 14. The correlation between dimension "cognitive" and dimension "support" was 0.78, which was a little bit higher that that of the international sample. The correlation between dimension "cognitive" and "interest" was again found to be very close to 0. And between the two attitude dimensions, the correlation was moderately high (r=0.47).

Table 14: Correlations among the Three Dimensions for Chinese and U.S. Students

	Cognitive Dimension	Interest Dimension
Interest Dimension	0.03	
Support Dimension	0.78	0.47

The reliabilities of dimensions "cognitive", "interest" and "support" were 0.83, 0.60, and 0.74, respectively. The reliability of dimension "interest" was again found to be the lowest among the three. It is higher than that for mainland Chinese and Hong Kong students, but lower than that for U.S. - Canadian sample. This is probably because that the items in this subscale do not work as well for two Asian groups as for U.S. and Canadian students.

The item infit statistics indicated that there were 8 items (4.2%) that did not fit the model's expectation very well. The deviance statistics reduced 14391.58 from the unidimensional model. And Chi-square was significance with 3 degrees of freedom.

When the MRCMLM was fitted to the data, a total of 60 items were found to exhibit substantial DIF. The number reduced by 36.8% (35 items) from the unidimensional approach. Table 15 presents the number of DIF items favoring each group found in the multidimensional analysis. Among the 60 items, 31 favored U.S. students, and 29 favored mainland Chinese students. 47 items (78.3% of the 60 DIF items) had effect sizes larger than 0.638 (Class C DIF), about half favoring each group of students. The largest item-by-group interaction reduced only a little bit to 1.70 logits.

Table 15: Multidimensional DIF Analysis between Chinese and U.S. Students

	Number of Class B DIF Items	Number of Class C DIF Items	Total Number of DIF Items
Favoring Chinese Students	6	23	29
Favoring U.S. Students	7	24	31
Total Number of Items	13	47	60

Multidimensionality again explained away all the DIF in the two attitude dimensions. 35 items, 22 in the "interest" dimension and 13 in the "support" dimension, which were found to display substantial DIF in the original unidimensional analysis, no longer had DIF under the multidimensional framework.

Systematic DIF

Although multidimensionality accounted for a big proportion of DIF, there remained 60 DIF items. This constituted 31.3% of the total number of items in the assessment, and made up over half of the "cognitive" items. Moreover, the majority of the DIF items had large effect sizes. Hence, it is crucial to investigate whether the test contains systematic DIF before we can use the test results to make any comparisons between U.S. and Chinese students.

We first compared the overall person proficiency estimates obtained from a model in which DIF items were excluded with those obtained from the original model in which all items in the assessment were included. The correlation was still very high (r=0.96), but lower than those found in the previous two comparison groups. The paired t-test is statistically significant, with a small effect size (d=0.15). Note that the difference between measures, though still small, was also larger than those found in the U.S.-Canada and mainland-Hong Kong analyses. Secondly, we compared the two person ability estimates of the "cognitive" dimension alone. The correlation dropped to 0.92. The t-test was again significant with a very small difference of 0.03. As t-tests were strongly influenced by sample sizes, the large sample size here can be the reason why the tests were statistically significant.

Because the effect size was quite small, it seems that DIF did not have a big impact for Chinese and U.S. students at test level. One possible reason why there was no systematic DIF with such a high proportion of DIF items might be that about half of the items favor U.S. students and half favor Chinese students.

As there was no test-level DIF, it seems to be safe to use their total scores to rank order the students, or make other comparisons or inferences. However, since the number of DIF items, especially the number of Class C DIF items, was appallingly large, the interpretations of the test scores, especially for the "cognitive" dimension, were likely to be qualitatively different for these two groups of students. Suggestions or inferences that are based on the use of specific item scores will not be appropriate.

In short, multidimensionality was to found be an important cause of DIF when the three distinct dimensions were analyzed from a unidimensional approach. A three-dimensional model should be applied to analyze the response data from U.S. and Chinese students. Under the multidimensional framework, the test results may be safely used to

make general overall comparisons. However, cross-country validity equivalence does not hold at item level for about half of the items.

Detailed content analysis was followed to explore potential causes of the remaining DIF, so that future test development efforts can better prevent those bias patterns, and more information at item level can, in turn, be utilized.

Causes of DIF between U.S. and Chinese Students

Test Translation The Chinese language, which is based on an ideographic writing system is radically different from the English language, which is based on an alphabetical system. When a test is translated from one language to another, the differences between the languages are most likely to alter the items in terms of vocabulary difficulty, sentence length, sentence structure, and contextual connotations (Ercikan, 1998). As reviewed in Chapter 3, the vocabulary difficulty may change after translation due to differential frequency of word use. Sentence length and sentence structure may change because grammatical forms either do not have equivalents, or else have many of them in one or the other language. And finally, the contextual meaning can be hard to convey across languages.

To examine whether these differences are probable causes of DIF in this study, the DIF items were examined with special attention to (a) vocabulary difficulty of the key words, (b) passage length, (c) grammatical structure of key sentences, and (d) passage contextual meaning.

After investigating the DIF items, I did not find any big differences in terms of vocabulary difficulty level, grammatical structure or passage contextual meaning. First, in terms of vocabulary difficulty, all key words seemed to be quite common in both languages. Secondly, the grammatical structures, although inevitably not equivalent in many places, do not contain any elements that are particularly difficult to one of the groups. Finally, passage contextual meaning did not seem to be a potential cause of DIF either. One possible reason of these is that all of the texts were expository. Expository texts explain things by definition, sequence, categorization, comparison-contrast, enumeration, process, problem-solution, description, or cause-effect. This kind of text uses facts and details, opinions and examples to inform and persuade. And understanding the texts usually does not require much effort to figure out any contextual meaning. (As opposed to narrative texts, which often include elements such as a theme, plot, conflict(s), resolution, characters, and a setting, and use story to inform and persuade.) (Burke, 2000)

However, passage length seemed to differ between the two language versions to a greater extent. In particular, items with long texts (more than 3 or 4 paragraphs) tend to be longer in English than in Chinese (counted in words). But although the English version tended to be longer in general, there were only 4 DIF items (6.8% of all DIF items) with noticeably longer texts. Further more, whether it required more reading time

and contributed to item difficulty level for U.S. students needs further exploration, but is beyond the scope of this research.

In summary, we found that only a few DIF items could probably be explained in terms of language differences. As discussed earlier, other differences between the comparison groups, such as cultural and curricular differences, could lead to DIF as well. In addition, the students and teachers we consulted from the two countries all reported that the reading load of the sample was moderate and students could understand the sample item with some effort. In general, when the texts were not very difficulty for examinees to understand, the impact of language differences on items' difficulty levels will be relatively small.

Curriculum Coverage Comparing the degree of content-curriculum match of the PISA 2006 Science assessment for 15-year-old students in U.S. and China is extremely challenging because the curricula differ considerably from state to state in U.S. (In China, most provinces use the same curricula.) Moreover, it is impossible to get feedback from U.S. students without IRB (Institutional Review Board) approval even though the consultation is anonymous and does not collect any identifiable information from the respondents. As a result, I only succeeded in collecting opinions from Science educators in the State of Georgia. Hence, the following discussions are actually comparisons of the content-curriculum match between Chinese and U.S. Georgian students.

Our result showed a differentiated test-curricula match between the two groups. Specifically, although all topics were reported to have been covered in class in both countries, the expectations for learning varied to some extent. Among the 29 DIF items that favored Chinese students, 10 were expected to be "mastered" by Chinese students, but only briefly introduced to U.S. students in Georgia. For example, Chinese teachers and students reported that students should have "master level" understanding of the necessary conditions needed for combustion, whereas U.S. students were only expected to have learned something about it but have not applied it in any complex ways. On the other hand, there were 5 DIF items that were found to be better covered in the curricula in Georgia. One example was that U.S. students were expected to be very good at using of key words on Internet search engine to look for information, while Chinese students were only expected to have some basic skills.

Furthermore, we once again noticed differential strength in transmitting "Knowledge About Science" (understanding of inquiry and the nature of science) and "Knowledge Of Science" (understanding of fundamental scientific concepts and facts) between the two groups. We found that most DIF items measuring "Knowledge About Science" favored U.S. students. Items on "how to design scientific experiments", "how to collect scientific evidences", "how to search library, Internet, etc to look for information" and "how to explain scientific phenomena" were more often than not easier for U.S. students than for Chinese students. And most of the DIF items favoring mainland Chinese students measure specific Science concepts and facts. These suggested that the curricula

in mainland China might put more emphases on specific science knowledge whereas the curricula in U.S. (at least in Georgia) might have stronger emphases on scientific inquiry. But this trend was not clearly shown from the consultation results. Again, the reason might be that the teachers and students I consulted have somewhat different understanding of the terms "master level" and "basic level" although they were defined at the beginning of the consultation. Or, it could be due to the lack of choices of levels of mastery.

In summary, differential curriculum coverage does seem to be an important cause of DIF between Chinese and U.S. students. We found that differentiated curriculum coverage was the direct cause of DIF for 15 items (25% out of the 60 DIF items). Furthermore, we suspect that "Know About Science" is better conveyed in U.S.'s curricula, while curricula in mainland China are better at transmitting more "Knowledge of Science" topics.

Cultural Difference Earlier, we suspected that differential social desirability and response style might be an important cause of DIF in attitude items. The quantitative DIF detection analysis revealed that under the multidimensional framework, none of the attitude items exhibited substantial DIF. However, there were indications of differential response style or differential social desirability between Chinese and U.S. students.

Specifically, the 22 items in the "interest" dimension that were found to exhibit DIF in the unidimensional analysis all favored Chinese students. In addition, the average item difficulties of all "interest" items were much lower for mainland students: when calibrated separately, the mean difficulty of this dimension was -0.04 for mainland students and 0.31 for U.S. students. This pattern might suggest an acquiescence response style (ARS, defined as a tendency to agree with items regardless of actual attitude) for mainland Chinese students or a disacquiescence response style (DARS, defined as a tendency to agree with items regardless of actual attitude) for U.S. students. Or, it could also be the result of differential social desirability: Chinese students might express more interest because of the expectation of their teachers, parents or the society. Of course, this could also reflect an actual difference in "interest" between the two groups of students.

In the "support" dimension, we found a reverse pattern. More DIF items detected in the unidimensional analysis favored U.S. students (8 out of 13). The average item difficulty of this dimension was -0.95 for U.S. students and -0.80 for Chinese students, suggesting that it was easier for U.S. students to show "support" than Chinese students. This might manifest a cultural difference: it is probably easier for U.S. students to show their support of scientific enquiry. Chinese students, on the other hand, tend to be more conservative in expressing their attitudes towards things that are not taught explicitly by their teachers and parents. And it is less likely for Chinese teachers to discuss "supporting scientific enquiry" as the instructional emphases lay more on specific facts and concepts. Nevertheless, this difference could, again, be the actual difference in students' attitude

towards "supporting scientific enquiry". What the true underlying reasons are for the difference awaits further exploration in the future.

Finally, we examined the DIF items in the "cognitive" dimension and found that differential content familiarity might explain some of the detected DIF. Specifically, we suspected that five items might contain contents that are more familiar to one group of students. For example, the item about the magnetic hover train again favored Chinese students. We were not surprised to find that the items on "grand canyon" favored U.S. students. In addition, items on "forest fire", "genetically modified food" and "sun screen" contained subjects that seem to be more familiar to U.S. students, and were found to favor U.S. students.

To sum it up, although there are indications of differential response style and differential social desirability, they are not important causes of DIF between Chinese and U.S. students in this test. However, differential content familiarity could be an important contributor of DIF between these two groups, causing about 10% of the DIF detected in the multidimensional analysis.

The detailed content analysis revealed that language difference might cause DIF in just a few items. Cultural difference, especially differential content familiarity can also lead to DIF between U.S. and Chinese students. Moreover, differential curriculum coverage seems to be a very important cause of DIF in the PISA 2006 Science assessment. There remain some DIF items that cannot be explained by any of three hypothesized causes. Future research may try to find more plausible explanations.

References

Buckley, J. (2009). *Cross-National Response Styles in International Educational Assessments: Evidence from PISA 2006.* Paper presented at the NCES PISA Research Conference.

Burke, J. (2000). *Reading Reminders: Tools, Tips, and Techniques*: Boynton/Cook.

Chinese Language (2010). 2010, from http://en.wikipedia.org/wiki/Chinese_language

DeFrancis, J. (1986). *The Chinese Language: Fact and Fantasy*: University of Hawaii Press.

Ercikan, K. (1998). Translation effects in international assessment. *International Journal of Educational Research, 29*, 543–553.

The Hong Kong education and schooling system explained (2005). 2010, from http://www.tuition.com.hk/education-system.htm

Wang, W.-C. (2008). Assessment of Differential Item Functioning. *Journal of Applied Measurement, 9*(4).

Wilson, M. (2005). *Constructing Measures: An Item Response Modeling Approach*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Chapter 5: Discussions

5.1 Synthesis of Results

In this "era of cross-cultural encounters"(Vijver & Poortinga, 1997), cross-national student achievement comparison studies have become more prevalent and more powerful than ever, offering researchers and policy makers information that cannot to be gained from single-system studies alone.

Test scores from international assessments are at the core of such studies. An implicit assumption of the use of the test scores is that the scores obtained by students from different nations or cultural groups have the same psychological meaning. That is, construct equivalence is an essential requirement for making valid cross-cultural comparisons. If the construct being assessed is not consistent across nations or groups, inferences based on the assessment results may be biased. In fact, the validity of all the inferences made based on international assessment results critically dependents upon the construct equivalence across countries (Ercikan, 1998).

As cross-cultural comparisons are becoming increasingly popular and important, the problem of item bias and its detection is receiving increased attention from test developers and researchers. Most efforts have been focused on test adaptation (Hambleton & Kanjee, 1995) during the test development process and post hoc statistical item bias detection (Millsap & Everson, 1993).

PISA, as one of the most important international student achievement assessments, has been concerned about the cross-country validity from the very beginning. Numerous efforts have been devoted to ensure the construct equivalence throughout the test construction process. However, whether the assessment is fair to all participating groups with vastly different cultural backgrounds remains a question. Previous research found moderate to high percentages of DIF between many comparison groups in the PISA 2003 assessments: for example, in the 2003 Science assessment, there were 52 DIF items (37%) in the Canadian comparison (English version vs. French version), 54 (39%) in the England–France comparison, and 110 (79%) in the United States–France comparison (Ercikan & Koh, 2005).

In this study, we investigated the validity equivalence of the PISA 2006 Science assessment between three comparison groups, namely, the Canadian and U.S comparison, the mainland Chinese and Hong Kong comparison, and the U.S. and mainland Chinese comparison. Through these analyses, we hope to capture the opportunity the PISA 2006 Science test presented to gain deeper insights into the possibilities and limitations of large-scale international tests to measure and report students' complex knowledge of science across languages and cultures.

In addition, we took one step further beyond the statistical DIF detection analysis and exerted great efforts to look for possible explanations of DIF via detailed content analyses where DIF were found to threaten the construct equivalence between the groups

of interest. Although statistical DIF detection procedures has been the core of item bias analysis so far, substantive analysis is of special importance as it brings up possibilities to eliminate bias in the future.

Findings from statistical DIF analyses

First, to answer the question whether the functional equivalence of the PISA 2006 Science scale between any of the three comparison groups was at risk, both unidimensional and multidimensional statistical DIF detection procedures were carried out for each comparison group. Table 16 summarized the results across all analyses.

Table 16: Summary of Statistical DIF Analysis Results

	Number of DIF items found in uni-dimensional analysis	Number of DIF items found in multi-dimensional analysis	Reduction in the number of DIF items from uni- to multi- dimensional analysis	Systematic test-level DIF
Canadian-U.S. students	41 (21.4%)	10 (5.2%)	31	no
Mainland Chinese-H.K. students	76 (39.6%)	36 (18.8%)	40	no
Mainland Chinese-U.S. students	95 (49.5%)	60 (31.3%)	35	no

A few interesting observations were made. First, looking across the lines of the table, we can see that the number (and percentage) of DIF items is the smallest between Canadian and U.S. students and the largest between U.S. and Chinese students. The difference between the proportions of DIF items across the three comparisons suggests that the degree the reference and focal groups differ in their social and cultural backgrounds can be a good predictor of the severances of DIF. In this study, Canadian and U.S. students have more similar languages, curricula and cultures and were found to have the fewest number of DIF items, whereas Chinese and U.S. students represent two groups that differ in almost every possible way in these aspects and had the most serious DIF problems. Moreover, this result also indicates that DIF is not a fixed character of any test item. It is not an inherent property of the test, but a function of the use and interpretation of the test scores.

Secondly, looking across the columns of the table, we noticed that for all three comparisons the number of DIF items reduced significantly when we analyzed the data using a multidimensional approach. The reduction of DIF from uni- to multi- dimensional approach confirms that when the items in a test measure more than one construct, whether intended constructs or unwanted nuisances, multidimensionality needs to be taken into account. Otherwise, results based on single-dimensional analysis can give rise

to DIF. In our case, the PISA science test was designed to measure three sub areas: the "cognitive" science knowledge, "support for scientific inquiry", and "interest in learning science". Hence, the design of the test is inherently multidimensional. The low correlations among the sub-areas also suggest that the test is empirically multidimensional. So a multidimensional model should be adopted. Further more, the low correlations between the attitude scales and the cognitive scale may suggest that the inclusion of "attitudes towards science" in the definition of scientific literacy requires further consideration.

Finally, although the percentages of DIF were quite high even under a multidimensional framework for two of the comparisons (the mainland Chinese-Hong Kong comparison, and the Chinese-U.S. comparison), none of the three pairs were found to have systematic DIF at test level. It is important to note that this does not imply that the test is bias-free. Instead, it is possible that item-level bias were distributed in such a way that the difference between two groups in the ability estimates is not affected. As a consequence, although it might be safe to use the average scores to rank order the overall performances of countries, we recommend that stakeholders use the test results with extreme caution. The data might not be suitable to provide detailed information to serve as a broad basis for school improvement decisions. Especially for groups with more different backgrounds, the large proportions of DIF are of greater concern. The interpretation of the science scale can be very unstable from one country to another due to those DIF items. Hence, at least for some countries, the important responsibility of PISA to ensure that the instrument provides reliable and fully comparable information to all other participating countries has yet to be fulfilled.

Findings from content analyses

In the past, the analysis of bias was often limited to the statistical detection of item bias. However, researchers have recently started to devote more efforts to identifying and understanding the sources of DIF in cross-cultural comparisons (Allalouf, Hambleton, & Sireci, 1999; Ercikan, 1998; Ercikan & Koh, 2005; K. Ercikan, 2002; K. Ercikan, Gierl, McCreith, Puhan, & Koh, 2004; Hambleton & Kanjee, 1995). Researchers found that bias can occur for a variety of reasons, including test translation, differential curriculum coverage, differential impact of item format, differential cultural appropriateness of item content, speediness and other physical test conditions, etc.

In our study, we conducted an integrated analysis in which three most plausible causes of DIF, language, curriculum and cultural differences, were scrutinized. The effects of each factor would be limited to one or a few items in the test, or it can also affect many items depending on the backgrounds of the comparison groups. Table 17 presents the number (and percentages) of DIF items associated with by each factor.

The results are self-explanatory: As Hong Kong and mainland Chinese students have more similar languages and cultures, the effects of language difference and cultural influence are minimal, and differential curriculum coverage appeared to be the most

significant cause of DIF. Whereas for U.S. and mainland Chinese students, language and culture differences explained about 15% of the DIF. Curriculum coverage was, again, found to be the most important cause, which explained 25% of the detected DIF.

Table 17: Summary of the Content Analyses

	Number of DIF items associated with language difference	Number of DIF items associated with curriculum difference	Number of DIF items associated with cultural difference	Number of DIF items not associated with any of the 3 factors
Mainland Chinese-H.K. students	0 (0.0%)	7 (19.4%)	1 (2.8%)	28 (77.8%)
Chinese-U.S. students	4 (6.8%)	15 (25%)	5 (8.3%)	36 (60%)

Among the three potential causes of DIF studies, language difference only accounted for a small proportion of DIF in the case where the two languages (English and Chinese) differ vastly linguistically. Part of the reason can probably be attributed to the earlier attempts of researchers to identify language factors affecting the DIF of translated items (Allalouf, et al., 1999; Hambleton & Patsula, 2000; OECD, 2006; Wolf, 1998) as well as the effort of test developers to address those factors (OECD, 2006). PISA exerted a great amount of effort to take those factors into account at an early stage in the test development process, thus resulting in one step closer to the final goal of construct equivalence.

PISA was also concerned about the impact of differential curriculum impact. Its innovative solution to the problem was to develop a "curriculum-free" measure of the overall "scientific literacy"(OECD, 2006). Unfortunately, contrary to the test developers' belief, differential curriculum coverage was found to be the most serious cause of DIF in both the Hong Kong-Mainland and the U.S.-Chinese comparisons.

Previous studies have also criticized the validity of cross-national achievement scores relating to the differing national curricula and the concomitant problems that arise in test development and reporting (Berliner, 1993; Linn & Baker, 1995; Westbury, 1993). Researchers and test developers have acknowledged that it was extremely difficult to identify what constitutes a "common" curriculum in science across countries. Given this, it is unlikely that PISA 2006 science assessment can equally represent the science curriculum of each participating country, and differential curriculum coverage will inevitably cause DIF. This unsatisfactory reality leaves researchers and science educators to think about a few important questions: What science knowledge and its processes should students have? What facts and concepts from physics, chemistry, biology, and the Earth sciences should be the basis for school science programs? What science should students know and be able to do as future members of workforces? And what the new generation of curriculum materials should include to help students develop the science abilities and skills required in the modern society?

Cultural difference is another widely discussed potential source of DIF (Buckley, 2009; Cheung, 1996; Vijver & Poortinga, 1997). As it is a very broad concept, we chose to look at three specifically aspects that were most plausible to cause DIF in this study, namely, social desirability, content familiarity, and response style. Among them, response style and social desirability were not found to cause DIF. Although there might be true differences in how students respond to attitude items or in their attitudes toward science as the average item difficulties of the two attitude subscales were lower for mainland students in both comparisons, these differences are referred to as differential impact, but not DIF.

The only aspect that was found to potentially cause DIF is differential content familiarity. The challenge of this source of bias is very similar to that of the curriculum coverage. It raises questions to science educators and researchers about what contexts can be the basis for introducing science and technology and how science might be taught relating to social issues such as health, environment, resources, energy efficiency, etc.

In summary, detailed content analysis is useful in understanding and identifying potential causes of DIF. The three analyses showed us that DIF can occur for many different reasons, including language difference, curriculum and cultural difference. But not any single factor is an inherent source of DIF. The legitimacy of any sources of DIF relies on the specific context of the cross-country comparison. The three sources of DIF studies here can explain less than half the detected DIF in each comparison. Further investigations of the sources of bias require the collection of additional data.

5.2 Significances and Limitations of the Study

Significances

The PISA 2006 tests were administered to about half a million students in over 50 countries. The results have received attention from the media around the world. The analyses of the PISA 2006 Science assessment results from the four countries and areas discussed above provided an adequate basis for a few recommendations about the appropriate use of the data: first, multidimensional model should be used to calibrate the response data, or the three sub-scales can be calibrated and discussed separately. Secondly, while it is appropriate to use the test results to compare Canadian and U.S. students' performance at both test and item level, it is not recommended to compare Hong Kong and mainland Chinese students or U.S. and Chinese students at item level. Policy makers need to be very cautious when making decisions regarding the curriculum, resources or pedagogy based on any direct comparisons using single items.

In addition, the detailed substantive analyses of this study yielded some suggestions for future international assessment development: while the language impact can be controlled through well-designed test translation and adaptation procedure, the

impact of differential curriculum coverage and content familiarity remain challenges to the cross-national validity equivalence.

While the conclusions of this study are of value to the research community interested in international science assessment, the findings are important from a methodological perspective as well. The quantitative as well as substantive analysis methods used in this study can identify both DIF items and the causes of DIF for many items in other cross-language cross-nation studies as well. Such identifications can improve the development of tests with multiple language versions and enhance their cross-cultural validity.

And finally, the results also raised some important questions regarding the new developments of science curricula and science education.

Major limitations

This study also suffered from some methodological limitations. First, although the validity check via DIF analysis is a useful way of determining the cross-country equivalence of international assessments, it can only be conducted after the tests have been administered. The DIF for a specific test cannot be anticipated or prevented before the test's administration. But we can learn from post-hoc studies. And the substantive analysis also relied on post hoc explanations concerning the presumed causes of DIF.

Another major limitation is that the substantive analysis cannot bring definitive explanation of DIF, but only suggests possible causes. This is partly due to the confidentiality requirement of the items. I could not gather a panel of bilingual experts to discuss the specific items. As a consequence, the investigations on the impact of language difference and cultural factors depend on my subjective judgments.

In addition, because of the very limited resources, I only managed to get a few responses to the consultations on the curriculum coverage from each group of interest. One major drawback of this limitation is that I was not able to account for the variation of curricula within each group. This may not be a problem for Hong Kong or mainland Chinese students as the curricula used within these two groups do not vary much. But the curricula used in different states or even school districts in U.S. vary considerably. Educational systems can vary even within countries, sometimes significantly. Moreover, different sub-national systems can reflect varying cultures, priorities, and goals that were not captured in this study.

And finally, the design of the consultation questionnaire itself has much room for improvement as well, for instance, the definitions of levels of mastery may be better defined.

5.3 Future Directions

Based on the findings of this study, a few further steps are possible. First, to deepen our understanding of the validity equivalence problem in large-scale international assessments, it is worthwhile to extend the DIF study to compare students from more countries and areas. Results from the three analyses in this study have limited generalizability. As more than 50 countries and areas participated in PISA 2006, we have the opportunity to investigate DIF between many different cultures. Findings from more analyses might suggest consistent or dissimilar patterns. Secondly, as PISA 2009 has already been administered, there is also possibility to conduct multi-year comparisons between the same groups to look for evidences of the consistency in the findings. Third, as one of the major limitations of this study is the imperfection of the curriculum coverage consultation, given more resources, it can be done at a larger scale to take into consideration the within country variations and to obtain more accurate information on the impact of curriculum.

Another direction of future research can focus on the development of new DIF detection methods, which can incorporate the multilevel nature of the data structure. PISA and other large-scale international assessments often collect student responses and other student and school level information. The data has a multilevel structure. Traditional DIF detection methods do not take into consideration the nested nature of the data, and may not be appropriate. As a result, the development of new multilevel DIF detection methods is called for. They may also lead to different conclusions on the validity equivalence of assessments.

Reference

Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the causes of DIF in translated verbal items. *Journal of Educational Measurement, 36*(3), 185–198.

Berliner, D. C. (1993). International comparisons of student achievement: A false guide for reform. *National Forum, XXII*(3), 25—29.

Buckley, J. (2009). *Cross-National Response Styles in International Educational Assessments: Evidence from PISA 2006.* Paper presented at the NCES PISA Research Conference.

Cheung, S. K. (1996). Reliability and factor structure of the Chinese version of the Depression Self-Rating Scale. *Educational and Psychological Measurement, 56,* 142-154.

Ercikan (1998). Translation effects in international assessment. *International Journal of Educational Research, 29,* 543–553.

Ercikan, & Koh (2005). Examining the Construct Comparability of the English and French Versions of TIMSS. *International Journal of Testing, 5*(1), 23 - 35.

Ercikan, K. (2002). Disentangling Sources of Differential Item Functioning in Multilanguage Assessments. *International Journal of Testing, 2*(3,4), 199 - 215.

Ercikan, K., Gierl, M. J., McCreith, T., Puhan, G., & 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), 301 - 321.

Hambleton, R. K., & Kanjee, A. (1995). Increasing the validity of cross-cultural assessments: Use of improved methods for test adaptations. *European Journal of Psychological Assessment, 11*(3), 147–157.

Hambleton, R. K., & Patsula, L. (2000). *Adapting Tests for Use in Multiple Languages and Cultures. Laboratory of Psychometric and Evaluative Research Report.*

Linn, R. L., & Baker, E. L. (1995). What Do International Assessments Imply for World-Class Standards? *Educational Evaluation and Policy Analysis, 17*(4), 405-418.

Millsap, R., & Everson, H. T. (1993). Methodology Review: Statistical Approaches for Assessing Measurement Bias. *Applied Psychological Measurement,, 17*(4), 297-334.

OECD (2006). *Assessing Scientific, Reading and Mathematical Literacy: a Framework for PISA 2006.*

Vijver, F. J. R. v. d., & Poortinga, Y. H. (1997). Towards an Integrated Analysis of Bias in Cross-Cultural Assessment. *European Journal of Psychological Assessment, 13*(1), 29–37.

Westbury, I. (1993). American and Japanese achievement...again. *Educational Researcher, 22*(3), 21—25.

Wolf, R. M. (1998). Validity issues in international assessments. *International Journal of Educational Research, 29,* 491—501.

Appendices

Appendix A: Consultation Questionnaire on U.S. Curriculum

Part 1

Please rate how well a typical 15-year-old student in your state (or area) has learned the following topics. Select from one of the three conditions:

1. Master level: The topic has been instructed in class, and students have deep understanding of the content, they also have had opportunities to use the knowledge in classroom discussions, solving problems in homework, etc.
2. Basic level: The topic has been introduced to students in class, but not emphasized. Students may not have had opportunity to learn all detailed aspects of the topic, or they have not had opportunities to use it for problem solving.
3. Not there yet: The topic has not been introduced to students in class yet. Students may not know anything about it.

Other comments are also welcome. Please write in the space given after the table.

Topics	1. Master level	2. Basic level	3. Not there yet
1. Sound wave travels (propagates) in a given direction.			
2. Change in volume when ice change into water.			
3. Photosynthesis			
4. Carbon dioxide emissions from energy sources.			
5. Evaporation makes the concentration of solutions higher.			
6. Gravity			
7. The amount of solar energy that can be harnessed depends on the time of the day.			
8. How plants spread their seeds.			
9. The law of buoyancy			
10. The volume of molecule, atom and electron			
11. Pollution can kill animals.			
12. The use of key words when using internet search engine.			
13. How to calculate rate.			
14. Antibiotics kills virus.			
15. Necessary conditions for combustion.			
16. How to extinguish fires.			
17. Very low temperature can prevent the growth of virus.			
18. The ocean influences the climate.			
19. The effect of temperature on metabolism.			
20. When a vertebrate is infected with a virus, antibodies are produced.			
21. Children and the elderly have weaker immune systems, so they are more likely to get infectious diseases.			

22. Car exhaust causes air pollution.	
23. Ways to prevent and control natural disasters.	
24. What a combustion reaction produces.	
25. Different materials have different thermal conductivity.	
26. Protein denaturizing	
27. The primary source of energy for human body is carbohydrates.	
28. The angle of sunlight changes at different times of the day.	
29. The Moon exhibits different phases as the relative geometry of the Sun, Earth and Moon changes.	
30. The difference between necessary condition and sufficient condition.	
31. Seasons differ in Southern and Northern Hemisphere.	
32. The function of human heart.	
33. Health benefits of exercise.	
34. How to read line graph.	
35. The law of electromagnet.	
36. The difference between physical and chemical reactions.	
37. How to read diagrams	
38. How fossils were formed	
39. Using scientific evidences to explain phenomena.	
40. The development of Science contributions to the development of modern society.	
41. How to design scientific experiments to compare the effects of different conditions.	
42. What is a scientific experiment?	
43. How to find scientific data/information (search library, internet, etc)	
44. Find evidences to support hypotheses.	

Other comments:

Part 2

Below is a sample Science item bundle. Please read it and rate how easy/difficult it is for a typical 15-year-old U.S. student to read and understand the text and questions. (No need to solve the problems.)

Consider the following newspaper report.

DUTCHMAN USES CORN AS FUEL

Auke Ferwerda's stove contains a few logs burning quietly with low flames. From a paper bag next to the stove he takes a handful of corn and puts it onto the flames. Immediately the fire flares up brightly. "Look here," Ferwerda says, "The window of the stove stays clean and transparent. Combustion is complete." Ferwerda talks about the fact that corn can be used as concerned, this is the future.

Ferwerda points out that corn, in the form of cattle food, is in fact a type of fuel too. Cows eat corn to get energy out of it. But, Ferwerda explains, the sale of corn for fuel instead of for cattle food might be much more profitable for farmers.

Ferwerda has become convinced that, in the long run, corn will be widely used as fuel. He imagines what it will be like harvesting, storing, drying and packing the grains in bags for sale.

Ferwerda is currently investigating whether the whole corn plant could be used as fuel, but this research has not been completed yet.

What Ferwerda also needs to consider is the amount of attention being focused on carbon dioxide. Carbon dioxide is regarded as the main cause of the increase of the Greenhouse effect. The increase of the Greenhouse effect is said to be the cause of the increasing average temperature of the Earth's atmosphere.

In Ferwerda's view, however, there is nothing wrong with carbon dioxide. On the contrary, he argues, plants absorb it and convert it into oxygen for human beings.

However, Ferwerda's plans may clash with those of the government, which is actually trying to reduce the emission of carbon dioxide. Ferwerda says, "There are many scientists who say that carbon dioxide is not the main cause of the Greenhouse effect."

Question 1:

Ferwerda compares corn used as fuel to corn used as food.
The first column of the table below contains a list of things that happen when corn burns. Do these things also happen when corn works as a fuel in an animal body? Circle Yes or No for each.

When corn burns:	Does this also happen when corn works as a fuel in an animal body?
Oxygen is consumed.	Yes / No
Carbon dioxide is produced.	Yes / No
Energy is produced.	Yes / No

Question 2:
In the article a conversion of carbon dioxide is described: "...plants absorb it and convert it into oxygen ...". There are more substances involved in this conversion than carbon dioxide and oxygen only.
The conversion can be represented in the following way:
carbon dioxide + water → oxygen + ( ) Write in the parentheses the name of the missing substance.

Question 3:
At the end of the article Ferwerda refers to scientists who say that carbon dioxide is not the main cause of the Greenhouse effect.
Karin finds the following table showing the relative Greenhouse effect caused by four gases:

Relative Greenhouse effect per molecule of gas			
Carbon dioxide	Methane	Nitrous oxide	Chlorofluorocarbons
1	30	160	17 000

From this table Karin cannot conclude which gas is the main cause of the increase of the Greenhouse effect. The data in the table need to be combined with other data for Karin to conclude which gas is the main cause of the increase of the Greenhouse effect.
Which other data does Karin need to collect?
A. Data about the origin of the four gases.
B. Data about the absorption of the four gases by plants.
C. Data about the size of each of the four types of molecules.
D. Data about the amounts of each of the four gases in the atmosphere.

Please rate how easy/difficult it is for a typical 15-year-old U.S. student to understand the text, and the question: (check one box below)
☐ easy to comprehend
☐ can comprehend most of it with some effort
☐ can only comprehend part of it with great effort
☐ very difficult to comprehend, cannot understand the text and/or the questions, cannot proceed to answer the questions as a result

Appendix B: Consultation Questionnaire on Hong Kong Curriculum

第一部分

請評判下面表格中所列出的內容，對於大部分香港15歲學生來說，哪些在課程中已深入教授 / 學習過，哪些簡要介紹 / 學習過，哪些從未涉及。請在對應的空格中打 “X”。

* 深入教授 / 學習: 指內容除了已由老師講解外，學生還運用該內容完成過課堂討論，課後習題等。

* 簡要介紹 / 學習: 指內容僅由老師講解過，學生隻大致了解，並沒有運用與討論、習題或其他方面。

如果有其他的看法，請在表格後的“其他意見”處寫明。

	1. 深入教授 / 學習	2. 簡要介紹 / 學習	3. 從未涉及
1. 水分蒸發對溶液濃度的影響			
2. 植物種子的傳播方式			
3. 密度與浮力的關係			
4. 物質彈性大小和形變的關係			
5. 分子，原子，和質子的體積大小			
6. 網絡搜索中關鍵詞的使用			
7. 峽谷形成的地質原因			
8. 速度的計算方法			
9. 滅火的原理			
10. 海洋對氣候的影響			
11.物理髮應與化學反應的區別			
12.燃燒的條件和產物			
13. 通過物質的分子式了解該物質的化學成分			
14.化石的形成過程			
15.人類能感知的聲音的頻率範圍			
16.電流的測量工具			
17.不同物質有不同的熱傳導性能			
18.人體能量的主要來源為碳水化合物			

19.太陽光照射角度在一天內的變化	
20.月球和地球處於不同相對位置時的不同月相	
21.百分比的計算	
22.人體心臟的功能	
23.抗生素使用的副作用	
24.運動對健康的好處	
25.運用科學道理的解釋現象的能力培養	
26.設計實驗，比較不同條件對結果的影響	
27.科學發展對現代社會的貢獻	
28.科學實驗的定義和作用	
29.充分條件和必要條件的區別	
30.如何查找可靠的科學數據 / 信息	

其他意見：

第二部分

下面是一道考試樣題。請閱讀題目中提供的短文和需要回答的問題，然後評價短文和問題對於大部分香港15歲學生來說是否易於讀懂。（不需要回答試題。）
請在題後對應的選擇項前打 “X”。

閱讀思考以下這一新聞報導。

荷蘭人利用玉米作燃料

奧克•菲爾達的爐灶下幾個原木燃著微火。從火爐旁邊的一個紙袋中，他手抓一把玉米然後和把它撒進火焰中。隨即，火苗變得很旺盛。“你看這裡”，菲爾達說，“火爐的四周保持清潔，火苗旺盛並且玉米能夠完全燃燒。”菲爾達說出了一個事實，即玉米可被用來作為燃料，就如同可以做牛的食物一樣。據他而言，這是不久後可以實現的。

菲爾達指出，玉米作為牛的食物，實際上是一種燃料。牛吃玉米，利用它以取得能源。但是，菲爾達解釋說，出售玉米為燃料而不是做為養牛的食物可能更使農民有利可圖。

菲爾達堅信，從長遠來看，玉米將被廣泛用作燃料。他幻想著，它收穫後貯存，烘乾和包裝，然後裝袋出售。

菲爾達目前正在調查是否整個玉米植株都可作為燃料, 但這項研究尚未完成。

菲爾達還需要考慮的問題集中於二氧化碳的排放量。二氧化碳被視為是導致溫室效應主要的原因之一。增加溫室效應使地球的平均氣溫增加。

在菲爾達看來，二氧化碳是沒有錯的。相反，他認為，植物吸收二氧化碳並轉化成氧氣為人類所用。

然而，菲爾達的計劃可能會與政府發生衝突，而這實際上是在試圖減少二氧化碳的排放。菲爾達說，“事實上有許多科學家認為，二氧化碳不是造成溫室效應的主要原因”。

問題1

菲爾達把以玉米作為食物和以玉米作為燃料做比較。

下表第一列為玉米燃燒時的情況。玉米在動物體內消化是否也與這相同？圈是或不是。

當玉米燃燒時：	玉米在動物體內消化是否也與這相同？
消耗氧氣	是/否
生產二氧化碳	是/否
產生能量	是/否

問題2 ：

在文章中形容轉換二氧化碳是:“……植物吸收二氧化碳並轉化成氧氣…… ”。除了二氧化碳和氧氣外，還有有更多的物質參與了這一轉化。轉化方程式為：

二氧化碳+水→氧+（　　）在括號中寫入相應的物質。

問題3：

在文章末尾，菲爾達說有許多科學家認為二氧化碳不是造成溫室效應的主要原因。卡琳在資料上查到導致溫室效應的主要為以下四種氣體：

導致溫室效應的氣體分子數			
二氧化碳	甲烷	氧化亞氮	氯氟烴
1	30	160	17 000

從這個表卡琳不能確定哪種是影響溫室效應的主要氣體。表中的數據不足以說明問題，結合其它數據卡琳才能得出是什麼氣體主要導致了溫室效應而還有什麼數據是卡琳應該收集的？

A. 關於這四種氣體的產源。

B. 植物所能吸收這些氣體的量。

C. 這四種氣體的分子式。

D. 這四種氣體在大氣中的總排放量。

請選擇：

☐ 短文和問題都很容易理解。

☐ 短文或問題讀起來略有難度，但還是可以基本理解。

☐ 短文或問題很難讀懂，只能理解部分內容。

☐ 短文或問題太難讀懂了，不能理解，根本沒法進一步回答試題。

Appendix C: Consultation Questionnaire on Mainland Chinese Curriculum

第一部分.

请评判下面表格中所列出的内容，对于大部分中国15岁学生来说，哪些在课程中已深入教授 / 学习过，哪些简要介绍 / 学习过，哪些从未涉及。请在对应的空格中打 "X"。

*** 深入教授 / 学习: 指内容除了已由老师讲解外，学生还运用该内容完成过课堂讨论，课后习题等。**

*** 简要介绍 / 学习: 指内容仅由老师讲解过，学生只大致了解，并没有运用与讨论、习题或其他方面。**

如果有其他的看法，请在表格后的"其他意见"处写明。

	1. 深入教授 / 学习	2. 简要介绍 / 学习	3. 从未涉及
1. 人类能感知的声音的频率范围			
2. 声音的传播方向			
3. 电流的测量工具			
4. 不同物质有不同的热传导性能			
5. 冰和水转变时的体积变化			
6. 光合作用			
7. 哪些能源的使用会释放二氧化碳			
8. 水分蒸发对溶液浓度的影响			
9. 重力			
10. 每天里不同时间能获得的太阳能不同			
11. 蛋白质的变性			
12. 人体能量的主要来源为碳水化合物			
13. 植物种子的传播方式			
14. 物质弹性大小和形变的关系			
15. 密度与浮力的关系			
16. 太阳光照射角度在一天内的变化			
17. 月球和地球处于不同相对位置时的不同			

月相	
18. 分子，原子，和电子的体积大小	
19. 环境恶化会造成生物数量减少或灭绝	
20. 网络搜索中关键词的使用	
21. 百分比的计算	
22. 充分条件和必要条件的区别	
23. 峡谷形成的地质原因	
24. 速度的计算方法	
25. 抗生素会杀灭细菌	
26. 燃烧的必要条件	
27. 灭火的原理	
28. 低温冷冻可以防止细菌生长	
29. 南、北半球的季节不同	
30. 海洋对气候的影响	
31. 人体心脏的功能	
32. 新陈代谢和温度的关系	
33. 细菌感染会使人体产生抗体	
34. 儿童和老人对疾病的提抗能力较差，所以更容易得传染病	
35. 抗生素使用的副作用	
36. 汽车尾气排放会造成大气污染	
37. 运动对健康的好处	
38. 如何读懂线形图表	
39. 电磁铁的特性	
40. 自然灾害的预防和控制	
41. 物理发应与化学反应的区别	
42. 燃烧的产物	
43. 通过物质的分子式了解该物质的化学成分	
44. 看懂示意图	

45. 化石的形成过程	
46. 设计实验，比较不同条件对结果的影响	
47. 运用科学道理解释现象的能力培养	
48. 科学实验的定义和作用	
49. 如何查找可靠的科学数据 / 信息	
50. 科学发展对现代社会的贡献	
51. 寻找支持假设的证据	

其他意见：

第二部分.

下面是一道考试样题。请阅读题目中提供的短文和需要回答的问题，然后评价短文和问题对于中国15岁学生来说是否易于读懂。（不需要回答试题。）

请在题后对应的选择项前打 "X"。

阅读思考以下这一新闻报道。

荷兰人利用玉米作燃料

奥克•菲尔达的炉灶下几个原木燃着微火。从火炉旁边的一个纸袋中，他手抓一把玉米然后和把它撒进火焰中。随即，火苗变得很旺盛。"你看这里"，菲尔达说，"火炉的四周保持清洁，火苗旺盛并且玉米能够完全燃烧。"菲尔达说出了一个事实，即玉米可被用来作为燃料，就如同可以做牛的食物一样。据他而言，这是不久后可以实现的。

菲尔达指出，玉米作为牛的食物，实际上是一种燃料。牛吃玉米，利用它以取得能源。但是，菲尔达解释说，出售玉米为燃料而不是做为养牛的食物可能更使农民有利可图。

菲尔达坚信，从长远来看，玉米将被广泛用作燃料。他幻想着，它收获后贮存，烘干和包装，然后装袋出售。

菲尔达目前正在调查是否整个玉米植株都可作为燃料, 但这项研究尚未完成。

菲尔达还需要考虑的问题集中于二氧化碳的排放量。二氧化碳被视为是导致温室效应主要的原因之一。增加温室效应使地球的平均气温增加。

在菲尔达看来，二氧化碳是没有错的。相反，他认为，植物吸收二氧化碳并转化成氧气为人类所用。

然而，菲尔达的计划可能会与政府发生冲突，而这实际上是在试图减少二氧化碳的排放。菲尔达说，"事实上有许多科学家认为，二氧化碳不是造成温室效应的主要原因"。

问题1

菲尔达把以玉米作为食物和以玉米作为燃料做比较。

下表第一列为玉米燃烧时的情况。玉米在动物体内消化是否也与这相同？圈是或不是。

当玉米燃烧时：	玉米在动物体内消化是否也与这相同？
消耗氧气	是/否
生产二氧化碳	是/否
产生能量	是/否

问题2 ：

在文章中形容转换二氧化碳是:“……植物吸收二氧化碳并转化成氧气…… ”。除了二氧化碳和氧气外，还有有更多的物质参与了这一转化。转化方程式为：

二氧化碳+水→氧+（　　）在括号中写入相应的物质。

问题3：

在文章末尾，菲尔达说有许多科学家认为二氧化碳不是造成温室效应的主要原因。卡琳在资料上查到导致温室效应的主要为以下四种气体：

导致温室效应的气体分子数			
二氧化碳	甲烷	氧化亚氮	氯氟烃
1	30	160	17 000

从这个表卡琳不能确定哪种是影响温室效应的主要气体。表中的数据不足以说明问题，结合其它数据卡琳才能得出是什么气体主要导致了温室效应而还有什么数据是卡琳应该收集的？

A. 关于这四种气体的产源。

B. 植物所能吸收这些气体的量。

C. 这四种气体的分子式。

D. 这四种气体在大气中的总排放量。

请选择：

☐ **短文和问题都很容易理解。**

☐ **短文或问题读起来略有难度，但还是可以基本理解。**

☐ **短文或问题很难读懂，只能理解部分内容。**

☐ **短文或问题太难读懂了，不能理解，根本没法进一步回答试题。**