The influences of home language, gender, and social class on mathematics literacy in France, Germany, Hong Kong, and the United States

Aminah Perkins, Laura Quaynor, and George Engelhard, Jr.
Division of Educational Studies, Emory University, Atlanta, Georgia, USA

The purpose of this study was to examine the influences of home language, gender, and social class on mathematical literacy within the context of four countries (France, Germany, Hong Kong, and the United States) whose students participated in the mathematics section of the 2003 Programme for International Student Assessment (PISA). Rasch (1980) measurement theory was used to examine the effects of the three independent variables on students' performance on the mathematics test items; particular attention was paid to home language. Use of differential group (DGF) and differential person functioning (DPF) provided additional detail regarding variation and aberrant responses on the test of 84 items related to mathematical literacy. Home language had a statistically significant effect in Germany and Hong Kong, but not in France and the United States. As expected, gender and social class had statistically significant effects in all four countries on mathematics literacy, with the exception of gender in France. The DGF and DPF analyses illustrated group and individual variations in item responses related to home language within France.

IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments Volume 4

INTRODUCTION

Large-scale international assessments provide rich data that allow researchers to explore the relationships among different variables within a variety of national contexts. Different countries serve as a social laboratory where scholars can explore the connections between variables such as gender, race, social class, and school achievement. In this study, we used data from the mathematics section of the Organisation for Economic Co-operation and Development (OECD)'s Programme for International Student Assessment (PISA) 2003 for four of the participating countries—France, Germany, Hong Kong, and the United States—in order to investigate relationships among home language, gender, social class, and mathematical literacy.

We were also interested in identifying individuals and groups of students within these four countries whose response patterns on PISA 2003 were dissimilar to those of other students with comparable levels of mathematical literacy. We therefore designed our study so that we could explore differential person functioning (DPF).[1] We particularly wanted to know if some students' mathematics literacy might not have been accurately assessed because their home language differed from the test language, and if patterns of performance on the mathematics items were differentiated according to student gender and social class. Furthermore, we sought to identify students whose person response functions (PRFs) were dissimilar to the PRFs of other students with similar achievement levels. In some instances, a student can be so unlike other examinees that his or her overall test score is not an appropriate representation of his or her mathematical literacy (Levine & Rubin, 1979).

When conducting our secondary analyses of the PISA 2003 data for the four countries, we investigated the following four research questions; the fourth provided the main focus of our study.

1. Is there a relationship between mathematics achievement and assessment of students in their home language?
2. Is there a relationship between mathematics achievement and gender?
3. Is there a relationship between mathematics achievement and social class?
4. Can we use person response functions to examine data-to-model fit for individuals and groups?

We also explored the interactions between home language, gender, and social class.

1 We explain this term and the related term "person response functions" on page 39–40.

RELEVANT LITERATURE

Mathematics Achievement and Home Language

Mathematics is the school subject that teachers and researchers tend to view as the subject most accessible to linguistic minority students (see, for example, Rolka, 2004). As such, we hypothesized that mathematics was the subject least likely to differentiate the performance of linguistic minority and linguistic majority students. Although not all linguistic minority students are immigrants and not all immigrant students belong to linguistic minorities, it is often the case that students who are immigrants are also in the linguistic minority. In this section, we discuss results from PISA testing that are relevant to both groups of students.

PISA reports unique patterns of achievement and immigrant status within different countries. For example, in France and Germany, students with an immigrant background are more likely than students without an immigrant background to perform at the lower levels of achievement, while in Australia and Macau (China), the two groups have similar levels of mathematical literacy (OECD, 2006). However, of more importance than the differences in academic performance between nonimmigrant and immigrant students is the limited academic performance of many of the latter group of students. In some countries, large proportions of both first- and second-generation immigrant students[2] do not achieve a basic level of mathematics proficiency. For example, only 40% of first-generation immigrant students in France and 25% of such students in Germany reached Level 2 on the PISA achievement scale—"a baseline level of mathematics proficiency ... at which students begin to demonstrate the kind of skills that enable them to actively use mathematics; for example they are able to use basic algorithms, formulae and procedures, to make literal interpretations and to apply direct reasoning" (OECD, 2006, p. 8). In Germany and the United States, only one-third of second-generation immigrant students reached Level 2, raising questions about the ability of schools not only to adequately prepare all students as mathematically literate adults but also to assimilate these children socially. Another possible explanation is that immigrants within the different countries vary in the extent of their previous academic attainment.

The OECD (2006) reports that student home language accounted for some of the differences in mathematics achievement on the PISA 2003 test between immigrant and non-immigrant students in some countries. After controlling for parents' educational and occupational status, the PISA researchers found that the performance gap associated with the language spoken at home remained significant in Belgium, Canada, Germany, Hong Kong, Macau, the Russian Federation, and the United States. Within this group, achievement differences between students who spoke the language of the test at home and students who did not ranged from 13 achievement scale score points in Canada through 46 points in the United States to 90 points in Germany (Miller, Sen, & Malley, 2007).

2 First-generation immigrant students are students who were born abroad. Second-generation students are students born in the country of residence but with one or both parents born abroad.

Mathematics Achievement and Gender

Many countries continue to report differences in mathematics achievement between male and female students, with the differences generally favoring males. This gap, however, has diminished in recent years (Liu & Wilson, 2009a; Mullis, Martin, Gonzalez, & Chrostowski, 2004; OECD, 2004). Of the 41 countries that participated in PISA 2003, 12 showed no significant gender differences in mathematics literacy, 27 revealed a gender difference in favor of males, and one (Iceland) reported a gender difference in favor of females (OECD, 2004). In some contexts within and across countries, no gender differences emerged between boys' and girls' responses to certain types of mathematics items, although gender differences remained with respect to items in other formats or focused on specific mathematical domains (OECD, 2004; Robertson, 2005).

Liu and Wilson (2009b) drew on PISA 2003 data to compare the mathematics performance of students in the United States and Hong Kong. They found no gender differences in the United States for multiple-choice items. In Hong Kong, however, males outperformed female students on the same items. In the United States, female students outperformed male students on items related to probability, but in Hong Kong, there was no statistically significant difference between male and female students on these same items. These performance differences related to gender were, however, small compared to the differences in mathematics achievement overall between students in the United States and students in Hong Kong.

Else-Quest, Hyde, and Linn (2010) showed that mathematics achievement on PISA and on the International Association for the Evaluation of Educational Achievement's (IEA's) Trends in International Mathematics and Science Study (TIMSS) was related to other contextual indicators of societal gender differences. Lower female mathematics achievement was associated with lower female representation in government, research, and economic activity. Similarly, van Langen, Bosker, and Dekkers (2006) reported that female participation in science, technology, engineering, and mathematics (STEM) careers was higher in countries where the performance of female students was relatively close to that of male students on PISA. The authors of the PISA 2003 report (OECD, 2004) suggested that student attitudes toward mathematics may also have been associated with performance on the PISA 2003 test.

Mathematics Achievement and Social Class

Although social class is a difficult variable to define across national contexts, PISA collects information about the occupations of students' parents that can help to serve as a proxy for social class. The International Labor Office (ILO) provides an international framework for occupations, which PISA researchers use to code the parental occupational status of students.[3] These occupational codes are then cross-referenced to the International Socio-Economic Index (ISEI) (Ganzeboom & Treiman, 1996).

3 If both parents are in paid employment, PISA uses the higher occupational status of the two as the index of social status.

Student performance in mathematics on PISA 2003 strongly related to parental occupational status. Each standard deviation of difference on the ISEI was associated with 34 points of difference on the PISA mathematics scale (OECD, 2004). Some countries reported larger differences in student achievement based on parental occupation than did others. In Belgium, France, Germany, Hungary, Luxembourg, the Slovak Republic, and Liechtenstein, students whose parents had the lowest-status jobs scored similarly to students in the lowest-performing countries (OECD, 2004). This relationship was mediated by between-school performance differences. Overall, students at schools serving a majority of students from lower socioeconomic status (SES) homes performed at much lower levels in mathematics than low-SES students attending schools where the majority of students came from high SES backgrounds (Carey, 2008; Marks, 2006; OECD, 2004).

Precision of Person Measurement

We used differential person functioning (DPF) to identify students and groups of students whose performance on the PISA 2003 mathematics items differed from the performance expected under the Rasch (1980) model of analysis. DPF occurs when an individual's observed response pattern differs from the expected response pattern for individuals with the same measured latent trait. When an individual's response pattern is unlike what would be expected given the model of analysis, we refer to his or her responses as "unexpected."

The notion of person invariance is not new and dates back to the work of early researchers such as Mosier (1940, 1941). Person fit, or the variability in person responses on a particular test, is of considerable importance: if individual students' response patterns are unexpected, then using their total score on a test to represent their mathematical literacy is likely to be misleading. This concern with respect to person reliability is a theme evident across the works of Keats (1967), Lumsden (1977, 1980), and Mosier (1940, 1941).

The functional relationship between the probability of a person giving a correct response and his or her actual achievement level can be graphically represented through the use of a person response function (PRF). Graphical depictions of this relationship can be traced back to the work of Weiss (1973) and Lumsden (1977). Weiss called his graphical representation of item difficulties and individual responses to items a "trace line": his proposal was that as items increase in difficulty, the percentage of responses that a person is likely to answer correctly decreases. Lumsden (1977, 1980), who used psychological measurement and mental growth as the backdrop for his work, provided a useful approach to addressing issues of person reliability. He introduced the use of the person characteristic curve (PCC), equivalent to what is referred to as a PRF in this study: "The person characteristic curve is the plot for a single subject of the proportion of items passed at different difficulty levels. It is perfectly analogous to the item characteristic curve" (Lumsden, 1977, p. 478).

Lumsden was the first researcher to clearly define this term, although the idea was implicit in previous research by Keats (1967), Vale and Weiss (1975), and Weiss (1973).

The underlying idea behind PRFs is that a person receives a correct response on an item when his or her location on a latent variable is greater than the given location of an item. PRFs are concerned with an individual's response to items representing various difficulty levels, in contrast to item response functions (IRFs), where the focus is on the response of individuals who, as a group, will exhibit different levels of achievement on one specific item (Carroll, 1983; Mosier, 1940, 1941).

Lumsden (1977, 1980) also identified issues associated with using a total test score for grouping individuals. For example, two students may receive the same total test score, but when their scores are examined in relation to their correct responses on items that are ordered by difficulty, their person response functions will cross. This crossing illustrates the impact of the individuals' different response patterns. As Lumsden (1977, p. 481) states, the crossing of PRFs results in the estimates of reliability being "biased by the difficulty of the items." For teachers, knowing which situations can lead to crossing of PRFs could help them determine the instructional strategies most suitable for individual students.

Performances exhibiting crossing PRFs can lead to problems with the substantive interpretation of person performance. The ordering of persons below and above the intersection points vary when PRFs cross (Perkins & Engelhard, 2009). If PRFs do not cross, then persons are ordered in the same way across item subsets, thereby achieving item-invariant measurement. Crossing PRFs, however, yield person ordering that varies as a function of the difficulty of the item subsets. Our study provides illustrations regarding the potential utility and substantive value of PRFs.

METHOD

The PISA Assessment

PISA, an international assessment jointly developed by the participating OECD countries, covers the domains of mathematics, reading, and science literacy. One third of the assessment items are in multiple-choice format, one third in closed constructed-response format, and one third in open constructed-response format. Results from PISA are reported as scale scores, with the scale having an average score of 500 and a standard deviation of 100. PISA uses a balanced incomplete block (BIB) design, which means that each student responds to a subset of the items, questions are shared across the subsets, and the Rasch model is used "to scale the student data to derive the various comparative measures that are produced and reported by the OECD" (Turner & Adams, 2007, p. 246).

The mathematics section of PISA was designed to evaluate mathematical literacy, defined by the OECD (2003, p. 15) as "an individual's capacity to identify and understand the role that mathematics plays in the world, to make well-founded judgments and to use and engage with mathematics in ways that meet the needs of that individual's life as a constructive, concerned and reflective citizen." PISA 2003 contained 84 mathematics items divided into four content domains (space and shape, quantity, uncertainty, and change and relationship) and described mathematics literacy on a scale comprising three competency clusters.

Because Rasch measurement models enable the development of measurement scales in the form of a line or a variable map, test items, such as those in PISA, can be placed along this line according to their levels of difficulty. Individuals can also be placed on this line according to their level of achievement. In Rasch measurement, this construction is referred to as a variable map. Figure 1 displays the scale for mathematics literacy as a variable map identifying the expected and hypothesized location of students and mathematics items.

Students participating in PISA also fill out a questionnaire that asks them to report personal and family characteristics. Students self-identify the language spoken at home, as well as such characteristics as parental occupation, educational level, and attitudes toward school (OECD, 2003). This information enabled us to determine whether or not the item difficulty locations were invariant across the language, gender, and social groups within the four countries that we considered.

Figure 1: Hypothesized variable map

Latent variable: mathematical literacy		
Logit scale	*Students*	*Item*
HIGH 5.00 4.00 3.00 	• Interpret more complex information and negotiate a number of processing steps	*Reflection* • Original mathematical approach • Multiple complex methods • Generalization
2.00 1.00 0.00 -1.00 	• Typically carry out more complex tasks involving more than a single processing step	*Connections* • Standard problem-solving, translation, and interpretation
-2.00 -3.00 -4.00 -5.00 LOW	• Typically carry out single-step processes that involve recognition of familiar contexts and mathematically well-formulated problems and reproduction of well-known mathematical facts or processes	*Reproduction* • Routine computations, procedures, and problem-solving

Participants

PISA 2003 was administered to students between 15.3 and 16.2 years of age. Representative samples of at least 4,500 students were selected from at least 150 schools within each of the 41 countries that participated in the mathematics assessment. Table 1 displays the demographic characteristics of the students from the four countries featured in our study. The number of students across the four countries was 18,894, with 4,300 in France, 4,660 in Germany, 4,778 in Hong Kong, and 5,456 in the United States. As is recommended for researchers conducting comparative analyses of data sets from various countries, we used senate weights to adjust the calibration weight of each country to the same sample size (M. von Davier, personal communication, December 25, 2010). In line with Rutkowski, Gonzalez, Joncas, and von Davier's (2010) recommendation for analyses of large-scale survey data, we used a weight of 1,000 students per country for all of the analyses presented here.

Students speaking languages at home other than the language in which the test was administered ranged from 6.5% in France to 8.6% in the United States. According to the data on parental occupation—the variable used as a measure of SES—Hong Kong had the smallest percentage of students from white-collar highly skilled families (27.0%); the United States had the largest. The highest percentage of students from blue-collar low-skilled families was found in Hong Kong (21.5%); the lowest in the United States (4.0%). It is important to emphasize that these students self-identified their parents' occupations, which were then cross-referenced in accordance with PISA practice to the aforementioned ISEI (Ganzeboom & Treiman, 1996). Although students self-identified their parents' occupations, the trends of parental occupations across countries seemed to pair with other data related to occupational status. For example, in Hong Kong in 2003, 20% of adults were working as managerial and professional staff (Legislative Council, 2007), while in the United States, sociologists estimate that roughly half of the population comprises white-collar workers (Beeghley, 2004; Thompson & Hickey, 2005).

Country Selection

We intentionally chose France, Germany, Hong Kong, and the United States as the focus of our analyses. We wanted to consider countries with very different immigration histories and overall levels of achievement on PISA. France, as a former colonial power, receives immigrants from former colonies and follows an educational strategy of assimilation (Castles, 2004). Germany is a European state that recruited immigrants for labor after World War II, and originally provided separate schools for the children of immigrants (Castles & Miller, 2003). Although often considered an emigration society (OECD, 2006), Hong Kong has received many immigrants from mainland China (Carroll, 2007) and lost emigrants to Australia, Canada, the United Kingdom, and the United States (Salaff & Siu-Lun, 1995). In contrast, the United States is a country that was formed on the basis of immigration (OECD, 2006). Overall, students from Hong Kong scored in the top third of countries on the PISA mathematics assessment, students from Germany and France scored in the middle third, and students from the United States scored in the bottom third (OECD, 2006).

Table 1: Student characteristics

	Country							
	France (*n* = 1,000)		*Germany* (*n* = 1,000)		*Hong Kong* (*n* = 1,000)		*United States* (*n* = 1,000)	
	N	%	*N*	%	*N*	%	*N*	%
Home language								
Test language	897	89.7	803	80.3	879	87.9	873	87.3
Other language/dialect	65	6.5	67	6.7	86	8.6	86	8.6
Missing	39	3.9	130	13.0	35	3.5	41	4.1
Gender								
Male	474	47.4	498	49.8	502	50.2	504	50.4
Female	526	52.6	491	49.1	498	49.8	496	49.6
Missing	0	0.0	10	1.0	0	0.0	0	0.0
Social class								
White-collar, high-skilled	717	51.7	469	46.9	270	27.0	641	64.1
White-collar, low-skilled	229	22.9	256	25.6	328	32.8	207	20.7
Blue-collar, high-skilled	111	11.1	112	11.2	158	15.8	52	5.2
Blue-collar, low-skilled	109	10.9	71	7.1	215	21.5	40	4.0
Missing	34	3.4	92	9.2	29	2.9	59	5.9
Mathematical literacy (overall)								
M	510.9	502.99	550.38	482.89				
SE	2.50	3.32	4.54	2.95				

Note: OECD average is 500.72 (*SE* = .63). Demographic information is based on weighted data using senate weights to obtain a fixed sample size of 1,000 in each country.

Theoretical Framework

Our study was essentially an exploration of item-invariant measurement of persons and groups. Engelhard (in press) describes five requirements of invariant measurement that must be met to yield useful inferences for measurement in the social, behavioral, and health sciences:

1. The measurement of persons must be independent of the particular items that happen to be used for the measuring: *item-invariant measurement of persons*.
2. A more able person must always have a better chance of success on any item than a less able person: *non-crossing person response functions*.
3. The calibration of the items must be independent of the particular persons used for calibration: *person-invariant calibration of test items*.
4. Any person must have a better chance of success on an easy item than on a more difficult item: *non-crossing item response functions*.
5. Items must measure a single underlying latent variable: *unidimensionality*.

Requirements 1 and 2 address issues related to PRFs. These two requirements define the major focus of this study.

Data Analysis

Rasch measurement theory informed the analyses of the data. We used the FACETS computer program (Linacre, 2010) to conduct this work.[4] Figure 2 illustrates the conceptual framework underpinning our analyses. Here we can see that the latent variable of interest is mathematical literacy, made observable through the 84 mathematics items included in the mathematics section of PISA 2003. The observed responses are both dichotomous and polytomous. Student characteristics, such as home language, gender, and social class, may influence student achievement levels; and they are included as potentially construct-irrelevant sources of variation in the model.

The following partial credit model illustrates the main effects of our conceptual model:

$$Ln\left[\frac{P_{nijmpkj1}}{P_{nijmpkj0}}\right] \theta_n - \delta_i - \alpha_j - \lambda_m - \gamma_p - \tau_{k_i} \qquad [1]$$

where

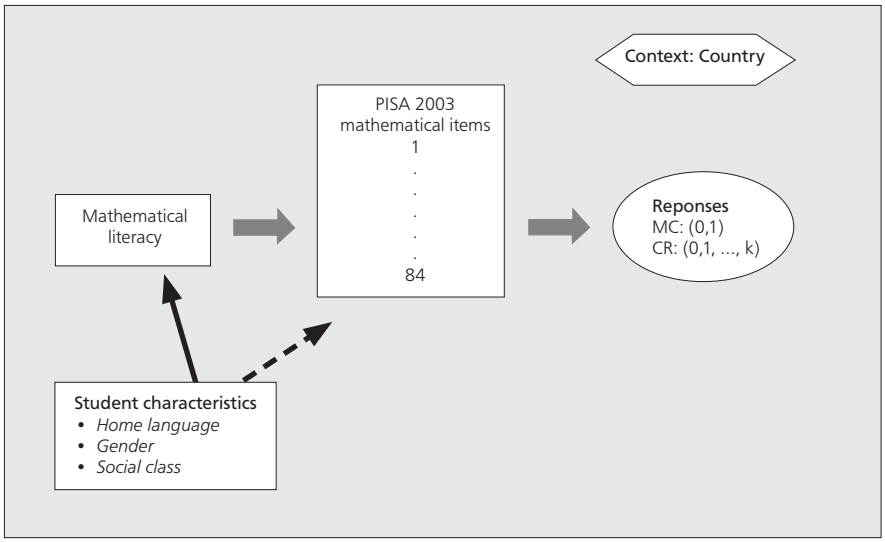
$P_{nijmpkj1}$ = the probability of person n succeeding on item *i* for group *j*, group *m*, group *p*, and threshold *k*,

$P_{nijmpkj0}$ = the probability of person *n* failing on item *i* for group *j*, group *m*, group *p*, and threshold *k*,

θ_n = the location of person *n* on the latent variable,

4 Note that the FACETS software uses joint maximum likelihood estimation. This estimation method is known to produce bias in parameter estimates (Andersen, 1972; Haberman, 1977). We applied the appropriate bias correction available in FACETS when conducting our analyses.

Figure 2: Conceptual model



δ_i = the difficulty of item *i*,

α_j = the location of language group *j*,

λ_m = the location of gender group *m*,

γ_p = the location of social class group *p*, and

τ_{ki} = the k^{th} threshold parameter of item *i*.

We examined main effects for the five facets—students, test items, home language, gender, and social class—across the four countries of interest. We subsequently used the main effects model to generate fit statistics, infit mean-square and outfit mean-square, as well as reliability of separation and chi-square statistics.

Infit mean-square and outfit mean-square are fit statistics that quantify the degree to which items or persons deviate from the expected model. The infit statistic is the sum of the squared-standardized residuals (Z^2_{ni}) summed over each element within a facet. This variance is then averaged by dividing it by the number of items the individual responded to, after which it is weighted by the individual's variance (W_{ni}) to account for the impact of the outliers, resulting in an infit statistic of the type seen in Equation 2 (Bond & Fox, 2007; Petridou & Williams, 2007). For this reason, infit is referred to as the information-weighted sum.

$$Inf\ it = \frac{\Sigma\, Z_{ni}^2 W_{ni}}{\Sigma W_{ni}} \quad [2]$$

The outfit statistic is calculated similarly, as seen in Equation 3. The difference between the infit and outfit statistics lies in the fact that the residuals are not weighted.

$$Ontf\ it = \frac{\Sigma\, Z_{ni}^2}{N} \quad [3]$$

Item difficulties were anchored at the actual item difficulties used in PISA 2003, based on the total number of participating countries. The student facet was not centered at zero. The other facets in the model were centered at zero (similar to contrast coding in multiple regression analyses) to estimate the contrasts between the elements within each facet.

RESULTS

As is apparent in Table 2, the reliability of separation for the items signifies that the PISA 2003 mathematics items included in our study manifested an array of difficulties. The chi-square statistic for the students and mathematics test items showed that the mean differences in students' mathematical literacy were statistically significant at the 0.05 level. Reliability of separation for gender and social class ranged from 0.63 to 0.89 across the countries. Gender and social class mean differences were also statistically significant in each of the four countries, except France, where gender had no statistically significant influence on mathematical literacy. The language facet had a statistically significant influence on mathematical literacy in all countries.

Figures 3 through 6 present the empirical variable maps for each country in the study. Each variable map displays the location of the five facets on the same scale of measurement. For example, Figure 3 shows that the male and female students sampled in France have the same location on the latent variable mathematical literacy. However, Figure 5 shows that male and female students sampled in Hong Kong have different locations on mathematical literacy, with females having a higher location on the latent variable. Readers should be cautioned that these are very small differences on the logit scale; one rule of thumb suggests differences less than 0.30 logits may not have substantive significance.

The variable maps also show a wide spread of item difficulties within each country. The percent of variance collectively explained by the five facets in the model is high at 76.9% in France, 76.0% in Germany, 79.1% in Hong Kong, and 74.5% in the United States. These are quite good degrees of model-data fit, and they provide support for the inference that the test is unidimensional.

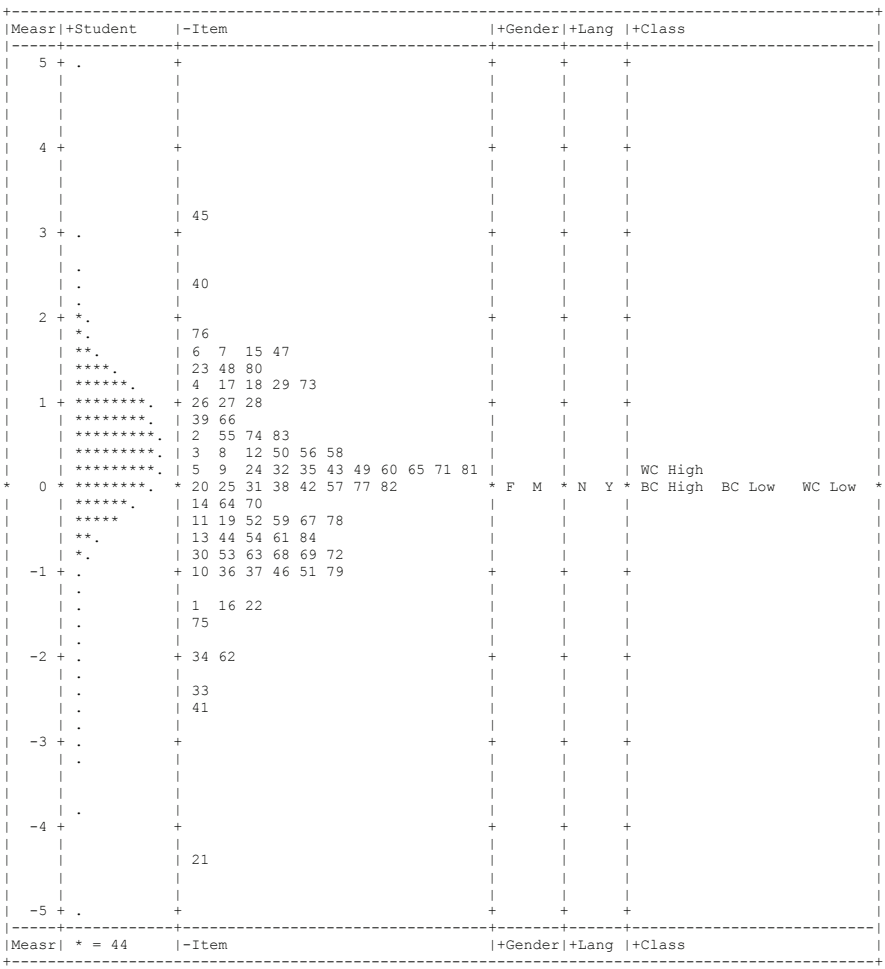
The expected value of the mean-square error statistics (infit and outfit) was 1.00, with a standard deviation of 0.20. The infit statistics for the facets across countries were all within the expected range of 0.80 to 1.20. However, the standard deviation of the infit for students was somewhat higher than expected, a finding that may reflect individual differences. This finding prompts additional explorations of person fit.

Table 2: Summary of FACETS statistics main effects

	France					Germany				
	Students	*Items*	*Language*	*Gender*	*Class*	*Students*	*Items*	*Language*	*Gender*	*Class*
Measures										
M	0.44	-0.01	0.00	0.00	0.00	0.42	0.18	0.00	0.00	0.00
SD	1.15	1.15	0.06	0.03	0.08	0.70	1.09	0.09	0.03	0.09
N	1,000	84	2	2	4	1,000	84	2	2	4
Infit										
M	1.05	1.02	1.07	1.08	1.09	1.05	1.05	1.14	1.09	1.10
SD	526	0.19	1.14	0.01	0.02	0.48	0.28	0.06	0.01	0.03
Outfit										
M	1.1	1.14	1.14	1.16	1.24	1.07	1.09	1.26	1.08	1.17
SD	0.94	1.0	1.49	0.01	0.15	0.86	0.72	0.22	0.22	
Reliability of separation		0.98	0.75	0.63	0.87		0.98	0.89	0.70	0.89
Chi-square statistic		6010*	8.0*	5.50	54.2*		1420.4*	18.5*	6.6*	54.5*
Degrees of freedom		83	1	1	3		83	1	1	3
France	**Hong Kong**					**United States**				
	Students	*Items*	*Language*	*Gender*	*Class*	*Students*	*Items*	*Language*	*Gender*	*Class*
Measures										
M	0.68	-0.01	0.00	0.00	0.00	-0.09	0.00	0.00	0.00	0.00
SD	0.76	1.07	0.08	0.03	0.05	1.13	1.07	0.05	0.03	0.09
N	1,000	84	2	2	4	1,000	84	2	2	4
Infit										
M	1.01	1.01	1.05	1.06	1.06	1.09	1.02	1.11	1.10	1.11
SD	0.49	0.18	0.03	0.01	0.02	0.46	0.24	0.01	0.01	0.02
Outfit										
M	1.10	1.12	1.31	1.18	1.18	1.05	1.04	1.08	1.05	1.08
SD	1.09	0.99	0.16	0.03	0.12	0.59	0.47	0.03	0.01	0.04
Reliability of separation		0.98	0.89	0.77	0.77		0.99	0.74	0.77	0.89
Chi-square statistic		6536.5*	17.6*	8.7*	18.4*		6725.1*	7.6*	8.9*	66.7*
Degrees of freedom		83	1	1	3		83	1	1	3

Note: *$p < .05$.

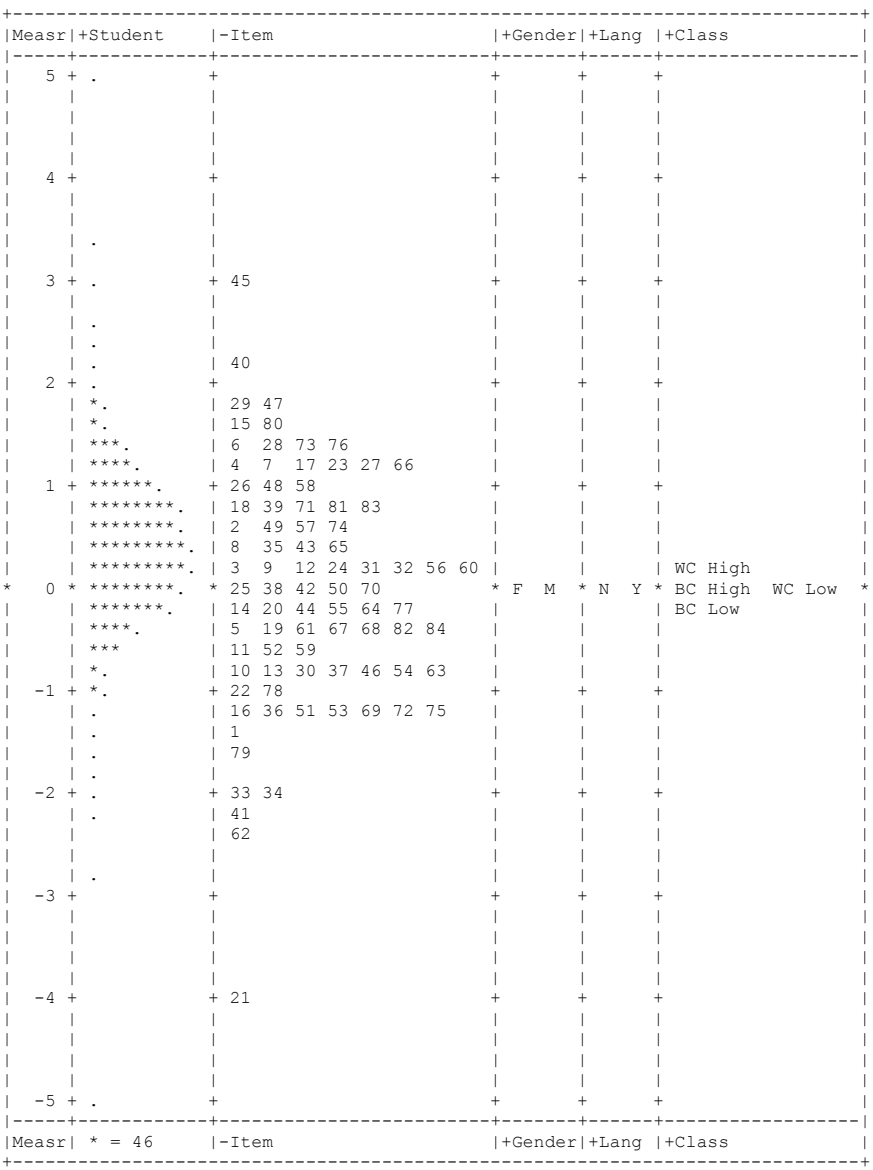
Figure 3: Empirical variable map for France



The outfit statistics for the facets across the countries were within the expected range, with the exception of the language facet. In both France and the United States, these values were higher than expected based on the model. The standard deviations of the outfit statistics were also high for the student and item facets across the countries.

Figure 7 shows the results of our exploration of group response functions. Here we can see the functional relationships between the probability of a correct response and the location of students within the two language groups. We used the location of the language groups—students who spoke the language of the test at home and students who did not speak the language of the test at home—along with the discrimination (slope) parameters to construct the group response functions for each country. Even though we found statistically significant differences for all countries, these displays

Figure 4: Empirical variable map for Germany



highlight the small substantive impact of these differences. Despite the overlap of language groups shown in Figure 7, there are still students within each group who exhibit unexpected response patterns.

Figure 5: Empirical variable map for Hong Kong

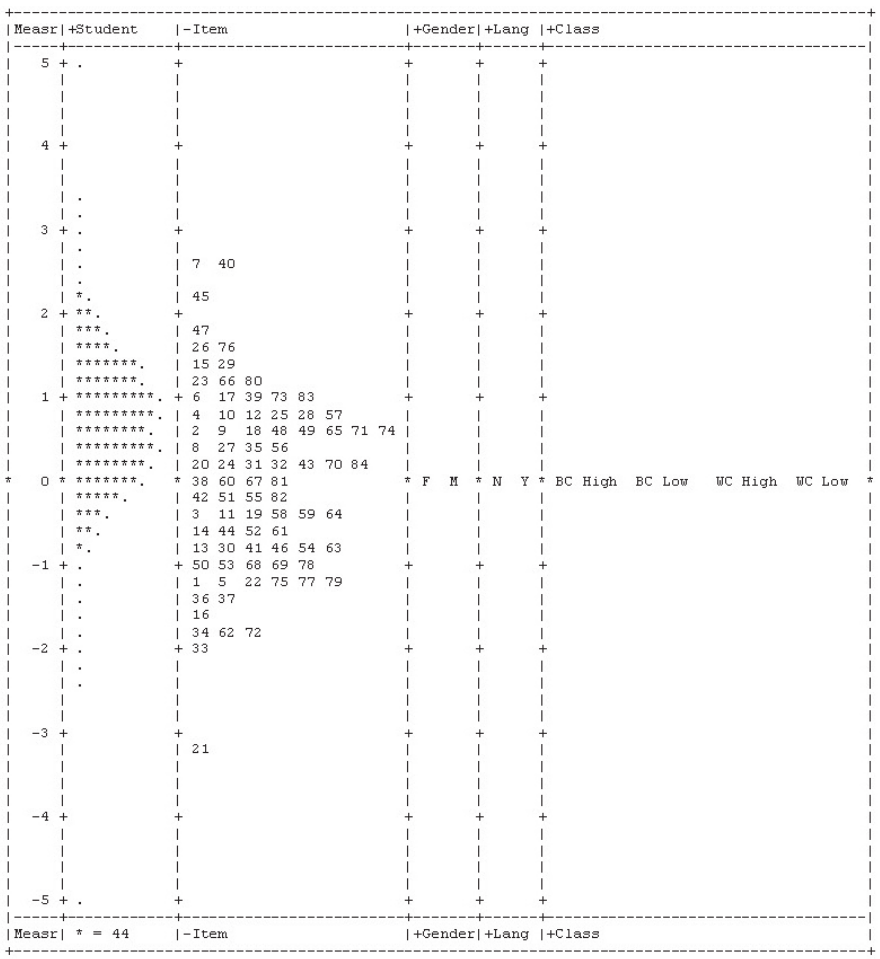


Figure 6: Empirical variable map for the United States

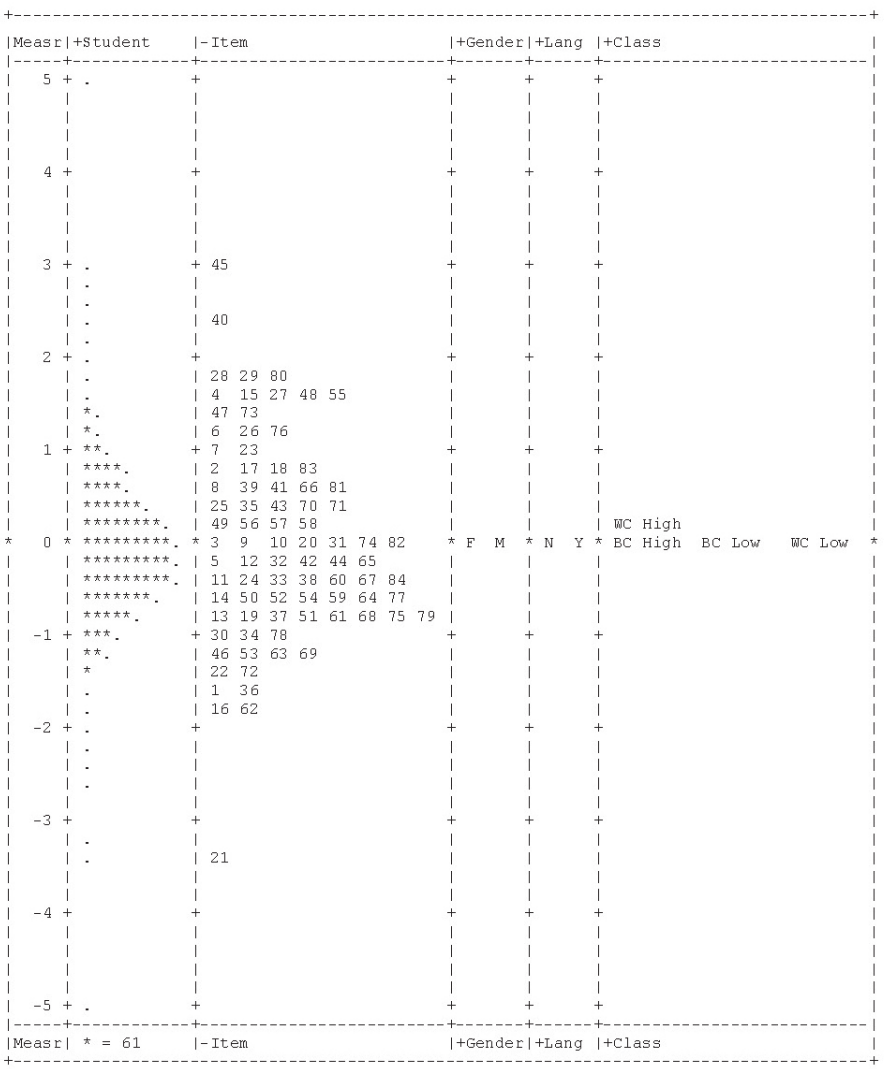
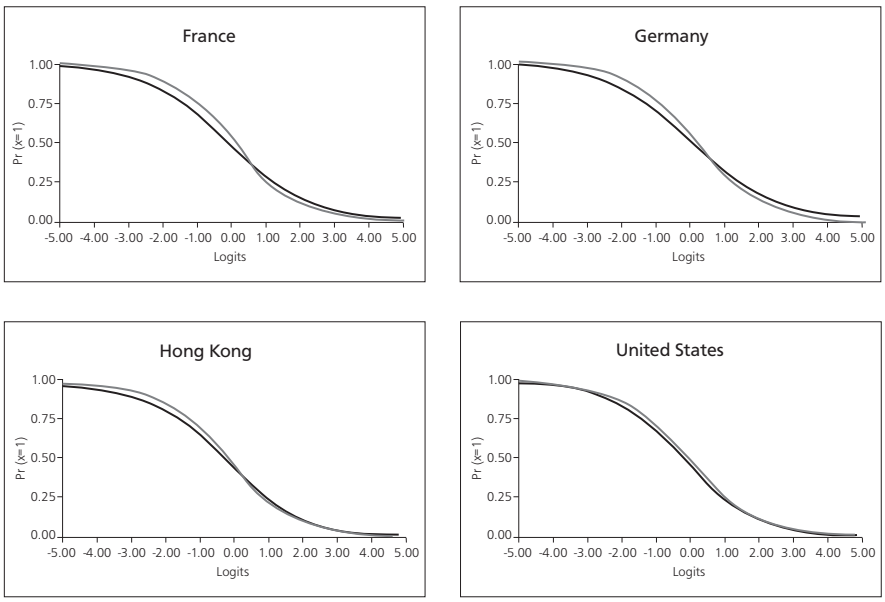


Figure 7: Language group response functions

	France		Germany		Hong Kong		United States	
	Lang (N)	*Lang (Y)*	*Lang (N)*	*Lang (Y)*	*Lang (N)*	*Lang (Y)*	*Lang (N)*	*Lang (Y)*
Location	-0.06	0.06	-0.09	0.09	-0.08	0.08	-0.05	0.05
Discrimination	0.81	0.99	0.78	1.00	0.86	1.01	0.90	0.97
Infit MNSQ	1.14	1.07	1.21	1.08	1.11	1.05	1.12	1.10
Outfit MNSQ	1.49	1.14	1.48	1.05	1.48	1.15	1.13	9.00

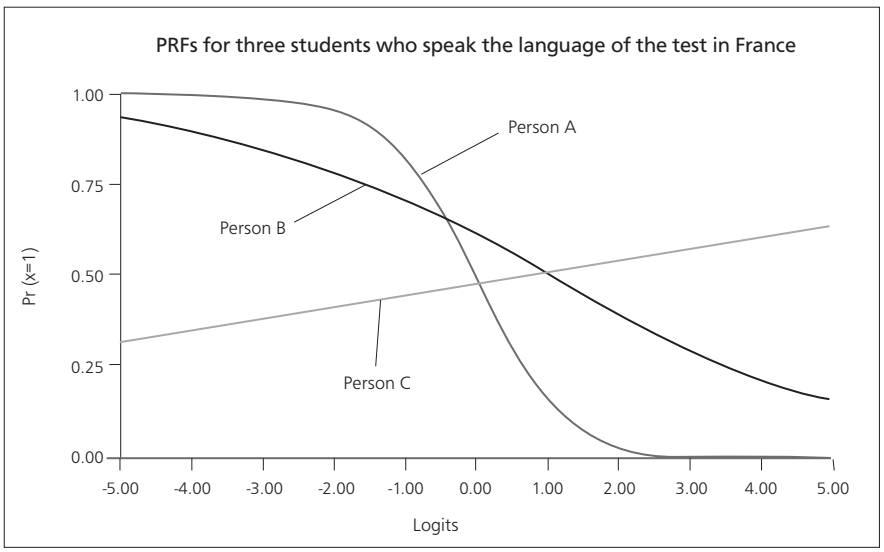


Note: All of the differences between language groups are statistically significant within countries ($p < .05$).

In order to illustrate these unexpected response patterns at the level of the individual, we chose three students from each language group in France and Germany; Figures 8 and 9 show the functional relationships for these individual-level student responses. The students we chose from each of these two countries had identical overall test scores (38 in France and 54 in Germany). In each language group, we identified a student in one of three fit classifications: less variation than expected (Persons A/D), fit the model (Persons B/E), did not fit the model (Persons C/F). Despite the statistically significant influence of the language facet, observations of the PRFs showed unexpected response patterns for certain students. The relationship between outfit and discrimination is also illustrated in the two figures. The correlation between these two student measures was fairly high at -0.649 in France and -0.711 in Germany.

Figure 8: Illustration of person response functions in France

	Language of test (Yes)			Language of test (No)		
	Person A	*Person B*	*Person C*	*Person D*	*Person E*	*Person F*
Location	-0.01	1.17	0.57	1.25	1.22	-0.25
Discrimination	1.60	0.42	-0.13	1.01	0.85	0.26
Infit MNSQ	0.48	2.26	1.59	0.41	0.67	1.47
Outfit MNSQ	0.48	1.18	1.63	0.39	1.06	1.52



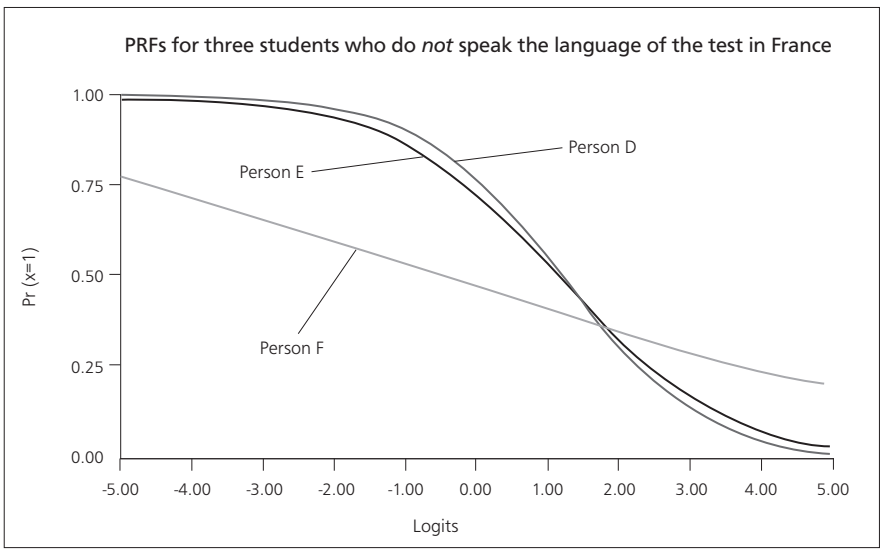
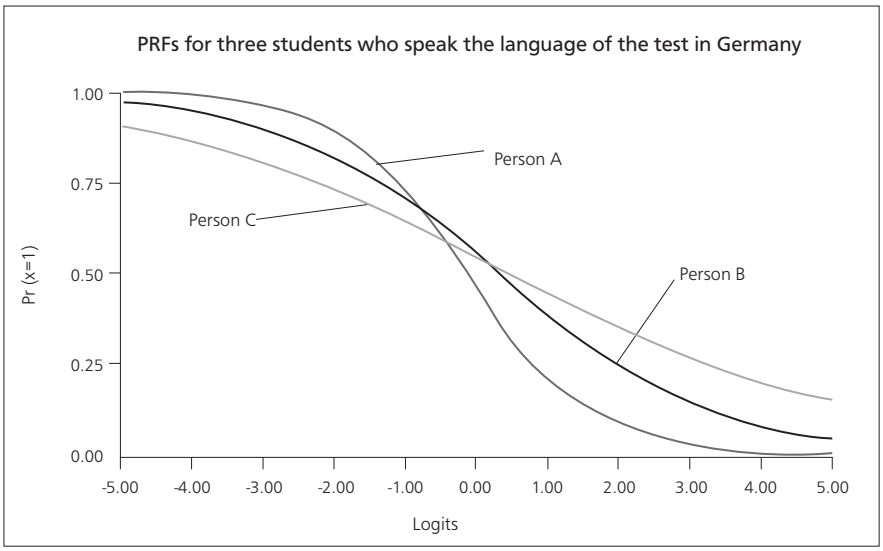
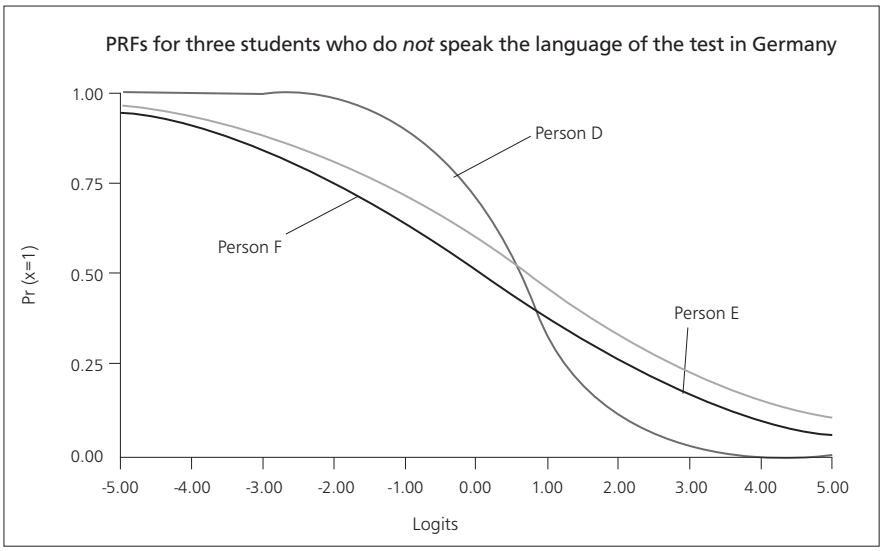


Figure 9: Illustration of person response functions in Germany

	Language of test (Yes)			Language of test (No)		
	Person A	*Person B*	*Person C*	*Person D*	*Person E*	*Person F*
Location	-0.18	0.35	0.50	0.60	0.05	0.76
Discrimination	1.12	0.64	0.40	1.42	0.54	0.51
Infit MNSQ	0.35	1.32	1.27	0.72	1.07	1.47
Outf i MNSQ	0.57	1.15	1.88	0.56	1.18	1.83





DISCUSSION

The focus of our study was exploration of the relationships between mathematical literacy and several student characteristics. Rasch measurement theory provided the methodological framework for conducting our analyses, which meant we could look at item-level rather than simply score-level information about students. Our analyses showed a relationship between home language and mathematics literacy within each of the four countries from which we drew samples of students who participated in PISA 2003 (i.e., France, Germany, Hong Kong, and the United States).

We found that gender and social class had significant influences on students' mathematics literacy in Germany, Hong Kong, and the United States, but not in France. Although we describe results in this paper as statistically significant, we do so with a caveat. The FACETS software that we used applies joint maximum likelihood (JML) estimation, which is known to produce biased parameter estimates (Andersen, 1972; Haberman, 1977). Although we applied the approximate bias correction available in FACETS to the analyses presented here, and although item difficulties were anchored on the actual values used in the PISA 2003 study, thus reducing the effects of using JML, we cannot be assured that the results reported as significant would be confirmed in a reanalysis using an unbiased estimation method.

The findings reported here are consistent with the findings of other recent research. Within an international context, gender differences in students' mathematics literacy continue to be salient across some countries. As we stated earlier, 28 out of the 41 PISA 2003 countries reported significant score-level gender differences (OECD, 2004). Social class, which we included in this study by using the students' parental occupation, is also a well-known predictor of achievement (OECD, 2006).

The use of group and person response functions offers a promising approach for examining aspects of student performance related to mathematical literacy on the PISA 2003 test items. The group response functions for language groups indicated that, despite the overall statistically significant effects for all countries, the small distances between the group response functions may not be substantively important. However, we still found significant individual differences in response patterns that may limit the inferences regarding mathematical literacy for some students. Figures 8 and 9 illustrate this phenomenon.

It is important to consider certain limitations that can also serve as recommendations for future research in the area of mathematics literacy. First of all, this study was a secondary data analysis. We had no direct control over the design of the assessment or decisions made with respect to the sampling design. As Rutkowski et al. (2010) point out, sampling weights should be used in order to adequately represent student achievement in each country. We used senate weights in our study to explore both national and individual levels of performance by students on each item representing mathematical literacy. Secondly, even though the main effects were statistically significant, the actual effect sizes on the logit scale were quite small. Because these small differences may not have any substantive significance, we again caution readers

against over-interpreting these main effects. Finally, there is a need to recognize the significant individual differences within each of the student subgroups examined here.

CONCLUSION

In summary, the results of this study on the influences of home language, gender, and social class on mathematical literacy confirm previous research that each of these factors is related to mathematical literacy. Our use of Rasch measurement provided results consistent with earlier findings based on PISA 2003. The study also illustrated how group and person response functions can be used to explore student response on mathematics items at a micro-level of analysis.

References

Andersen E. B. (1972). The numerical solution of a set of conditional estimation equations. *Journal of the Royal Statistical Society*, Series B, *34*, 42–54.

Beeghley, L. (2004). *The structure of social stratification in the United State*s. Boston, MA: Pearson, Allyn & Bacon.

Bond, T. G, & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Carey, D. (2008). Improving education outcomes in Germany. *OECD Economics Department Working Papers*, *No. 611*. Paris, France: OECD Publications.

Carroll, J. B. (1983). Psychometric theory and language testing. In J. W. Oller, Jr. (Ed.), *Issues in language testing research* (pp. 80–107). Rowley, MA: Newbury House Publishers.

Carroll, J. M. (2007). *A concise history of Hong Kong*. Lanham, MD: Rowman & Littlefield.

Castles, S. (2004). Migration, citizenship, and education. In J. A. Banks (Ed.), *Diversity and citizenship education: Global perspectives* (pp. 17–48). San Francisco, CA: Jossey-Bass.

Castles, S., & Miller, M. J. (2003). *The age of migration: International population movements in the modern world* (3rd ed.). New York, NY: Guilford.

Else-Quest, N., Hyde, J., & Linn, M. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin*, *136*, 103–127.

Engelhard, G. (in press). *Invariant measurement: Using Rasch models in the social, behavioral, and health science*s. New York, NY: Routledge.

Ganzeboom, H. B. G., & Treiman, D. J. (1996). Internationally comparable measures of occupational status for the 1988 International Standard Classification of Occupations. *Social Science Research*, *25*, 201–239.

Haberman S. J. (1977). Maximum likelihood estimates in exponential response models. *The Annals of Statistics*, *5*, 815–841.

Keats, J. A. (1967). Test theory. *Annual Review of Psychology*, *18*, 217–238.

Legislative Council. (2007). *Analysis of income disparity in Hong Kong*. Hong Kong, SAR: Legislative Council. Retrieved from http://www.legco.gov.hk/yr06-07/english/fc/fc/papers/fc0301fc-46-e.pdf

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics*, *4*, 269–290.

Linacre, J. M. (2010). *FACETS* (Version 3) [Computer program]. Chicago, IL: MESA Press.

Liu, O., & Wilson, M. (2009a). Gender differences in large-scale math assessments: PISA trends 2000 and 2003. *Applied Measurement in Education*, *22*(2), 164–184.

Liu, O., & Wilson, M. (2009b). Gender differences and similarities in PISA 2003 mathematics: A comparison between the United States and Hong Kong. *International Journal of Testing*, *9*(1), 20–40.

Lumsden, J. (1977). Person reliability. *Applied Psychological Measurement*, *1*(4), 477–482.

Lumsden, J. (1980). Variations on a theme by Thurstone. *Applied Psychological Measurement*, *4*(1), 1–7.

Marks, G. (2006). Are between- and within-school differences in student performance largely due to socio-economic background? Evidence from 30 countries. *Educational Research*, *48*(1), 21–40.

Miller, D. C., Sen, A., & Malley, L. B. (2007). *Comparative indicators of education in the United States and other G8 countries: 2006*. Retrieved from http://www.nces.ed.gov

Mosier, C. I. (1940). Psychophysics and mental test theory: Fundamental postulates and elementary theorems. *Psychological Review*, *47*, 355–366.

Mosier, C. I. (1941). Psychophysics and mental test theory. II. The constant process. *Psychological Review*, *48*, 235–249.

Mullis, I., Martin, M., Gonzalez, E., & Chrostowski, S. (2004). *TIMSS 2003 international mathematics report: Findings from IEA's Trends in International Mathematics and Science Study at the fourth and eighth grades*. Chestnut Hill, MA: Boston College.

Organisation for Economic Co-operation and Development (OECD). (2003). *The PISA 2003 assessment framework*. Paris, France: OECD Publications.

Organisation for Economic Co-operation and Development (OECD). (2004). *Learning for tomorrow's world: First results from PISA 2003*. Paris, France: OECD Publications.

Organisation for Economic Co-operation and Development (OECD). (2006). *Where immigrant students succeed: A comparative review of performance and engagement in PISA 2003*. Paris, France: OECD Publications.

Perkins, A., & Engelhard, G. (2009). Crossing person response functions. *Rasch Measurement Transactions*, *23*(1), 1183–1184.

Petridou, A., & Williams, J. (2007). Accounting for aberrant test response patterns using multilevel models. *Journal of Educational Measurement*, *44*(3), 227–247.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* (expanded edition). Chicago, IL: University of Chicago Press (Original work published 1960).

Robertson, I. (2005). Issues relating to curriculum, policy and gender raised by national and international surveys of achievement in mathematics. *Assessment in Education: Principles, Policy and Practice*, *12*(3), 217–236.

Rolka, K. (2004). Bilingual lessons and mathematical world views: A German perspective. In M. J. Hoines & A. B. Fuglestad (Eds.), *Proceedings of the 28th Conference of the International Group for the Psychology of Mathematics Education* (pp. 105–112). Cape Town, South Africa: International Group for the Psychology of Mathematics Education.

Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher*, *39*(2), 142–151.

Salaff, J., & Siu-Lun, W. (1995). Exiting Hong Kong: Social class experiences and the adjustment to 1997. In Ronald Skeldon (Ed.), *Emigration from Hong Kong: Trends and tendencies* (pp. 179–211). Hong Kong, SAR: Chinese University Press.

Thompson, W., & Hickey, J. (2005). *Society in focus*. Boston, MA: Pearson, Allyn & Bacon.

Turner, R., & Adams, R. J. (2007). The Programme for International Student Assessment: An overview. *Journal of Applied Measurement*, *8*(3), 237–248.

Vale, C. D., & Weiss, D. J. (1975). *A study of computer-administered adaptive testing* (Report 75-4, NTIS No. Ad-A018758). Minneapolis, MN: Psychometric Methods Program, Department of Psychology, University of Minnesota.

van Langen, A., Bosker, R., & Dekkers, H. (2006). Exploring cross-national differences in gender gaps in education. *Educational Research and Evaluation*, *12*(2), 155–177.

Weiss, D. J. (1973). *The stratified adaptive computerized ability test* (Report 73-3, NTIS No. AD-768376). Minneapolis, MN: Psychometric Methods Program, Department of Psychology, University of Minnesota.