Eğitim ve Bilim
2014, Cilt 39, Sayı 172
Geniş Ölçekli Test Uygulamaları Özel Sayısı

Education and Science
2014, Vol. 39, No 172
Large-Scale Assessment Special Issue

Profile Analysis as a Generalized Differential Item Functioning Analysis Method

Genelleştirilmiş Farklı İşleyen Madde Analiz Yöntemi Olarak Profil Analizi

Hüseyin H. YILDIRIM[1] **Selda YILDIRIM**[2]
Abant İzzet Baysal University

Norman VERHELST[3]
Eurometrics

Abstract

This study investigates the utility of a recent generalized differential item functioning analysis method based on profile analysis (PADIF). Two characteristics of PADIF are that the analysis is not at the item level but at the level of item categories and that the method can be used with an arbitrary number of groups. In the study, 169213 students' responses to the mathematics tests of the Programme for International Student Assessment (PISA) 2003 are used. The outcomes are aggregated at the country level. The results of PADIF are further evaluated in line with the findings in the literature. The results support PADIF as a promising method to reveal valuable information in international large scale assessments.

Keywords: profile analysis; Rasch analysis; differential item functioning analysis; PISA-2003 mathematics test

Öz

Bu çalışma, bir testte farklı işleyen soruları belirlemek üzere, profil analizine dayanan yeni bir metodu (PADIF) tanıtmakta ve bir örnek üzerinde işe yararlılığını incelemektedir. Bu genelleştirilmiş metodun klasik yöntemlerden iki önemli farkı analizlerin soru değil soru grupları üzerinde yürütülmesi ve sonuçların belirli bir öğrenci grubundaki bütünsel etkisinin görülebileceği şekilde bir araya getirilebilmesidir. Çalışmada 169213 öğrenciden elde edilen Uluslararası Öğrenci Başarısını Belirleme Programı (PISA) 2003 Matematik Test verisi kullanılmıştır. Sonuçlar ülkeler bazında bir araya getirilmiştir. Elde edilen bulgular bu alandaki literatürle karşılaştırılarak incelenmiştir. Sonuçlar PADIF'in uluslararası çalışmalarda gözden kaçan değerli bilgileri açığa çıkarmada ümit verici yeni bir metot olabileceğini göstermektedir.

Anahtar Sözcükler: Profil analizi; Rasch analizi; farklı işleyen madde analizi; PISA-2003 matematik testi

[1] Assist. Prof. Dr., Abant İzzet Baysal University, Faculty of Education, Golkoy, BOLU, Turkey; yildirim.huseyin@ibu.edu.tr

[2] Assist. Prof. Dr., Abant İzzet Baysal University, Faculty of Education, Golkoy, BOLU, Turkey; cet_s@ibu.edu.tr

[3] Norman Verhelst, PhD., Eurometrics, Tiel, The Netherlands; norman.verhelst@gmail.com

Introduction

In international large-scale assessments, such as the Programme for International Student Assessment (PISA), Differential Item Functioning (DIF) analyses are usually conducted to understand whether the different language versions of the tests are fair and equivalent across the participating countries. If researchers detect some items functioning differently across the countries, they then usually carry on their study to reveal the possible causes of DIF in those items detected (Yıldırım & Yıldırım, 2011). This kind of studies may contribute to the understanding of possible effects of item characteristics (e.g., item format, item content) or contextual variables (e.g., teaching practices, socioeconomic status) on responses of individuals (e.g., Klieme & Baumert, 2001).

From this perspective, DIF analyses can be considered as a supplementary technique to provide additional information to that of the measurement model (Zumbo, 2007). For instance, the measurement model applied to the PISA mathematics test is a generalized form of the Rasch model (OECD, 2005). In this model, test items are allowed to vary only in their difficulties, and the construct to be measured, mathematical literacy, is considered as being unidimensional. The model fits reasonably well the PISA mathematics test data and this allows parsimonious description of results so that a large non-specialized audience could access them. Nevertheless, it is known that this is not a perfect fit (e.g., Klieme & Baumert, 2001). That is, some aspects in the data are probably not covered by the measurement model.

For example, students in a country would probably perform relatively better on items they are familiar with through their learning experiences. If, say, graphical representations are of special importance in a country's teaching practices, one can expect students of that country to perform relatively better on mathematics items including graphics. However, the unidimensional Rasch model does not have a parameter to account for such country specific aspects. The model provides just a difficulty index to characterize an item across all participant countries. In other words, the measurement model cannot account for some country specific strengths (or weaknesses) due to differences in countries' teaching practices, learning experiences etc. In this context, DIF analysis can be regarded as a supplementary analysis to reveal such important aspects in the data which are missed by the measurement model.

However, to make use of DIF analysis in this manner researchers have to face two basic difficulties: 1) Identifying possible sources of DIF and 2) detecting the overall effect of DIF at a specific level of concern, such as the country level. The essential way in identifying sources of DIF has been exposing the DIF items (i.e., items flagged through DIF analysis as functioning differently across groups) to curriculum specialists' or test developers' interpretation. However, it is not uncommon that the judges fail to interpret most of the flagged items (Camilli & Shepard, 1994). According to Roussos and Stout (1996) the reason for this failure would be the lack of power because of the exploratory nature of a single-item analysis. Since judges focus on a single DIF item, the information they acquire would be too limited to identify possible sources of DIF.

The second difficulty lies in the fact that overall effect of DIF is usually investigated through a scale level analysis. However, research shows that item level DIF might not manifest itself in scale level analyses (Pae & Park, 2006; see Arim & Ercikan in this issue). For example, Zumbo (2003) conducted a comprehensive simulation study to measure the effect of item-level DIF on the test level. More specifically, he simulated data for two test groups based on a 3-parameter logistic Item Response Theory (IRT) model. Despite the fact that DIF was modeled on a specified number of items, the results of multi-group confirmatory factor analysis showed invariant factor structure across the groups. One of the reasons that DIF items did not manifest themselves at the test level was possibly that DIF at various items were cancelling each other at the test level (Pae & Park, 2006).

A recently introduced method by Verhelst (2012) based on Profile Analysis (PA) provides two promising advantages in dealing with these two difficulties mentioned above, and thus, may contribute to get the full benefit of DIF analysis as a supplementary analysis. This method can be regarded as a kind of generalized DIF and hereafter will be referred to as PADIF. Technical details on this method are provided in the next section.

Two characteristics of PADIF in dealing with the two difficulties in DIF analysis can be summarized as follows. First, as compared to classical DIF methods, analysis in PADIF is not carried on at the item level but at the level of sets or categories of items. This may contribute identifying sources of DIF because it is highly possible that sources of DIF may be more apparent in sets of items that share some characteristics. In this context, PADIF resembles the differential bundle functioning technique (Oshima, Raju, Flowers, & Slinde, 1998).

Second, to evaluate the overall effect of DIF (for example, at the country level), the method investigates the percentages of individuals in a country with respect to the gap between their observed and expected performances on a subset of test items. That is, the aggregation of the outcomes to the country level does not depend on the scale level analysis. Therefore, the problem of a possible DIF cancellation at the scale level is not an issue in PADIF.

In this context, the main objective of this study is investigating the utility of PADIF as a supplementary method that can provide some additional information to that of the measurement model. For this purpose, responses of individuals to the PISA-2003 mathematics test are used. The measurement model used in PISA is the Rasch model. The information revealed by PADIF analysis is further evaluated through a literature review to verify if the PADIF results are supported by some other research. PADIF analyses are carried out with respect to item formats, item lengths, contexts in which the items were presented, item contents, and competency levels items required. Further information on these categorizations is provided later in the text.

Method

PADIF Analysis

The analysis consists of two components: a profile analysis (PA) and an analysis aimed at showing differential item functioning. The former analysis explores if there are systematic differences between the test data and predictions following from the measurement model used. This analysis is carried out at the student level. The second component, DIF analysis, explores whether some kinds of discrepancies occur more or less systematically in predefined groups of test takers. In the framework of the PISA tests, these groups are defined as students from the same country. Both analyses are sketched in turn.

Profile analysis

The essence of the profile analysis lays in comparing the expected and observed performance of individuals on specific subsets of test items. To be more concrete, details are provided on a simplified example. Assume that a test is administered to a number of students and item parameters are estimated via the Rasch model which is shown to fit the test data. Also assume that items of a test are partitioned into two categories with respect to a criterion. For example, items that include a graphical representation are grouped in category A and the rest of the items are grouped in category B. In this case the main concern of profile analysis is comparing the expected and observed performance of individuals on category A and category B items.

Describing the observed performance of an individual is straightforward. Suppose that the total test score of a student is 6, and that this student obtained 4 on the items of category A and 2 on the items of category B. This is indicated as the ordered pair (4, 2) and called the *observed profile* of this student. Notice that the sum of the scores in the observed profile equals the total score.

The expected performance of an individual is specified as the *conditional expected profile* or *expected profile* for short, and it is the key concept in the profile analysis. The most important aspect of the expected profile is that it is calculated by using the item parameters estimates in the measurement model, which is the Rasch model in this example.

As is widely known the Rasch model can be written in two equivalent ways,

$$P(X_i = 1 \mid \theta) = \frac{\exp(\theta - \beta_i)}{1 + \exp(\theta - \beta_i)}$$

where $\xi = \exp(\theta)$ and $\varepsilon_i = \exp(-\beta_i)$ in which difficulty parameter of the *i*'th item is β_i.

To compute expected profiles one needs the so called basic symmetric functions of the ε-parameters

$$\gamma_s(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_k) = \sum_{*} \prod_{i=1}^{k} \varepsilon_i^{x_i}$$

In this equation *s* indicates the total test score of the individual for whom the expected score is being calculated, *k* is the number of test items and x_i is either 0 or 1. The '*' under the summation sign means that all possible response patterns which lead to a total score of *s* should be taken into account. These symmetric functions are used in detecting the conditional distribution of the response patterns given the test score. As revealed by G. Rasch (1960) this conditional distribution is dependent only on the item parameters. Finally, one can calculate through these symmetric functions, for example, that our hypothetical student who scored 6 on the test would be expected to score, say, 4.406 on category A items and 1.594 on category B items. This example is summarized in Table 1. Software, which can be requested from the third author, is available to conduct these calculations.

Table 1.
Example of Observed, Expected and Deviation Profiles

	Category *A*	Category *B*	Sum
Observed profile	4	2	6
Expected profile	4.406	1.594	6
Deviation profile	-0.406	+0.406	0

Notice that the sum of the two expected scores is equal (by definition) to the total test score. The difference between observed and expected profile is called the *deviation profile* (bottom row of Table 1). The sum of the deviations is zero by definition.

It is important to notice that the expected profile is calculated with respect to the item parameters estimated from the responses of all students who took the test. Thus, these parameters can be regarded as an indicator of the average performance of all the students. Consequently, the deviation profile in this case signals the individual differences.

For example the deviation profile in Table 1 shows that category B items are relatively easier for this student than estimated by the measurement model; because he scored better than expected on these items.

There are interesting questions to be asked in relation to Table 1. Focussing on a single student, one might ask for example if such a deviation from the expected profile is serious enough to pay special attention to category *A* items in the instruction of this student (by remedial teaching, for example) or if such a deviation is quite 'normal' and should not entail special actions (see Verhelst, 2012). On the other hand, in a large sample international survey one can also focus on groups of students, for example, all sampled students of a participating country to see if a deviation profile like the one displayed in Table 1 is prominent in a particular country, more than in all the countries or more than in another country. The answer to this question is the DIF part of the PADIF analysis.

DIF analysis

The first decision one has to take is about the use of what is meant by ‘similar’ deviation profiles. A simple typology would be this: either the student performs better than expected on category A items, or he performs better than expected on category B items (as in Table 1). (The probability that he performs equally well is virtually zero, as the expected scores are in general non-integer numbers.) The former type of student will be denoted as ‘A+’ and the latter as ‘B+’. In this case the student presented in Table 1 is of the type ‘B+’.

To do statistical analyses, one has to know what the probability is of an ‘*A*+’ or a ‘*B*+’ *type*, given the total score. To this purpose all possible observable profiles, compatible with the total observed score, should be considered along with their corresponding probabilities. These probabilities are also calculated in the same manner as the expected scores, by using the item parameters of the measurement model and several basic symmetric functions. For the imaginary case of a total score of 6, an example is given in Table 2. The *type* of each case is determined by comparing the observed profile to the expected profile (4.406, 1.594) specified above.

Table 2.
Probabilities of all Possible Observable Profiles Compatible with a Total Score of 6

Possible Observable Profiles	Type	Probability
(0,6)	*B*+	<0.001
(1,5)	*B*+	0.002
(2,4)	*B*+	0.026
(3,3)	*B*+	0.141
(4,2)	*B*+	0.348
(5,1)	*A*+	0.361
(6,0)	*A*+	0.122

From Table 2, it is easily deduced that the probability of being classified as an ‘*A*+’ type when the total test score is 6, is 0.361 + 0.122 = 0.483 (the sum of the two probabilities in the last two rows). This means, symbolically that $P(A+|S=6) = 0.483$, where S symbolizes the total test score, and trivially, that $P(B+|S=6) = 1 - P(A+|S=6) = 0.517$.

Once these probabilities are calculated, for the student exemplified in Table 1, an elementary contingency table can be built as the one given in Table 3

Table 3.
Elementary Contingency Table Based on a Single Observation

	Type *A*+	Type *B*+	Total
Observed frequency	0	1	1
Expected frequency	0.483	0.517	1

Of course, such a table is not very interesting if it is built only for an individual. But it is possible to construct such a table for all students from a country and then add them together. In Table 4 an actual example is given based on such an analysis conducted in this study for two countries (Russia – RUS and Turkey – TUR) participating in the PISA-2003 test for Mathematics. The item categories ‘Multiple Choice’ and ‘Constructed Response’ refer to item formats.

Table 4.
Contingency Tables for RUS and TUR Based on the PISA Mathematics Test of 2003

		Multiple Choice +	Constructed Response +	Total	E.P.(Multiple Choice)
RUS	Obs.	1571	1994	3565	-5.5
	Exp.	1767.1	1797.9	3565	
TUR	Obs.	1608	1209	2817	+7.2
	Exp.	1405.2	1411.8	2817	

Here is how to interpret such a contingency table. For Russia, the measurement model predicts that on the average 1767.1 students should exhibit a profile where they do better than predicted on the category 'Multiple Choice' and therefore worse than expected on the category 'Constructed Response'. But in the actual sample only 1571 did better on 'Multiple Choice' items. Apparently, the 'Multiple Choice' items are more difficult for the Russian students than predicted by the measurement model, and the category 'Constructed Response' is easier than predicted. To test if this tendency is statistically significant, a simple chi-square test with one degree of freedom can be applied to the Table for Russia.

For Turkey, just the opposite effect is observed: There are 1608 students of the 'Multiple Choice +' type (performing better on 'Multiple Choice' items than predicted), while the expectation is to find only 1405.2 such students. So, in Turkey the 'Multiple Choice' items appear to be easier than predicted by the model, just the opposite effect as found in Russia.

Joining the results of the two contingency tables in Table 4, this means that the test items function differently in the two countries considered. Since the partition of the items is based on a meaningful categorization, we are entitled to conclude that multiple choice format items are relatively easy in Turkey and relatively difficult in Russia, a conclusion which is along the lines of the conclusions drawn in DIF analyses. The difference with common DIF analysis is of course that the present analysis is not based on single items, but on item categories as a whole. Besides, this categorization gives an idea on the possible source of DIF, which is not the case in the analyses based on single items.

For reporting purposes, it may be useful to have a short summary of the contingency tables as given in Table 4. The *Excess Percentage* (EP) (reported in the rightmost column of Table 4) of a certain type is defined as the difference between observed and expected frequency of that type, expressed as a percentage of the sample size. Thus, EP(Multiple Choice) is 100 x (1571 – 1767.1)/3565 = -5.5 for Russia, saying that there are about 6% less students of the type 'Multiple Choice +' in Russia than predicted. If there are only two types, the excess percentage of the other category has the same absolute value but the opposite algebraic sign (i.e., EP (Constructed Response) = 5.5). In Turkey, the excess percentage for the same category is +7.2, saying that about 7% of the students more than predicted are of the type 'Multiple Choice +'. The chi-square test applied on the contingency table can be used to decide whether the EP differs significantly from zero, the value predicted by the measurement model.

Finally one should note that profile analysis can be conducted with more than two categories. For example constructed response format items in PISA consist of short answer format (i.e., the items that require students to construct a numeric answer, a single word or a short phrase) and extended answer (i.e., the items that require an extensive writing) format items. Thus, it is also possible to partition the mathematics items into three categories (i.e., multiple choice, short answer and extended response format items) for the profile analysis. One should notice that in such a case six different *types* can be distinguished such as (Multiple Choice +, Short Answer –, Extended Response –) or (Multiple Choice +, Short Answer +, Extended Response –). In this example a student of the former *type* performs better than expected on multiple choice format items but worse than expected on short answer and extended response format items, and a student of the later *type* performs better than expected on multiple choice format and short answer format items but worse than expected on extended response format items.

As it is clear in this example when there are more than two categories interpreting the analysis results gets complicated. Thus, for the ease of interpretation, in this current study, analyses of categorizations with more than two categories were dealt with in multiple sub-analyses in each of which one of the categories were analysed with respect to the rest of the categories as combined into a single category.

PADIF on the PISA-2003 Mathematics Test

The data

In PISA 2003 a rotated test design is used to produce 13 booklets. Each booklet consists of four clusters of items. The clusters contain either mathematics (the major domain) items, science items, reading items or problem solving items. The total number of clusters consisting of mathematics items is seven. During the test administration, one of these 13 booklets has been randomly assigned to each student. Forty-one countries participated in the PISA 2003 survey (OECD, 2005).

The data used in this current study consist of students' responses to the mathematics items in their booklets. Only the booklets which contain at least two clusters of mathematics items are selected for the analysis. These are the first six booklets which contain three clusters of mathematics items, and the 7th, 11th and 13th booklets with two clusters of mathematics items each.

The mathematics clusters contained in each of the selected booklets and the total numbers of mathematics items in each booklet are given in Table 5. The booklets selected for the analysis include all of the 84 mathematics items used in PISA 2003.

Table 5.

PISA Booklets Used in the Present Study

Booklets	Mathematics Clusters in Booklets	Number of Math Items
1	M1, M2, M4	36
2	M2, M3, M5	36
3	M3, M4, M6	35
4	M4, M5, M7	37
5	M1, M5, M6	37
6	M2, M6, M7	36
7	M3, M7	23
11	M1, M7	24
13	M1, M3	23

In this study responses of students who answered one of the nine booklets given in Table 5 are used. However a number of these students have been excluded from the analyses for two reasons. The first reason is that in the PISA database three different codes are used to code non-responses: not administered, not-reached and omits. A non-administered item is an item that has been excluded from the analyses in some countries because of, for example, severe translation errors. Such exclusions apply to all students in the country, but such an item might well be used in other countries. A sequence of items with no response that appears at the end of the test is coded as 'not-reached'. This sequence may vary in length from student to student. Other items with no response are coded as omits. In the PADIF analyses reported in the present study, students having more than five not-administered codes or not-reached codes (jointly) are excluded from the analyses. If a student is included, omits are treated as wrong answers. In the analyses the exclusion of not-reached and not-administered items is taken into account. For example: if a student made a test booklet having 35 mathematics items, but he has four not-reached codes, the analysis is performed as if the student has had that booklet not containing the last four items.

The second reason is that a profile is basically a distribution of correct and incorrect responses across item categories. If the total number of correct responses is very low or very high, the profile is either trivial or not very informative. To avoid inclusions of such profiles, students having less than three errors or less than three correct responses in total have been excluded from the analyses.

Table 6 presents for each country the number of students that has been involved in the analyses together with the percentage that these students represent to the total number of students who answered one of the nine booklets given in Table 5.

Table 6.
Number of Students Involved in the PADIF Analyses

Country	Acronym	Total	Percent	Country	Acronym	Total	Percent
Australia	AUS	8201	94.08	Korea	KOR	3582	95.19
Austria	AUT	3025	95.55	Liechtenstein	LIE	215	93.48
Belgium	BEL	5491	93.07	Luxembourg	LUX	2567	93.35
Brazil	BRA	1944	63.10	Latvia	LVA	2916	90.87
Canada	CAN	18334	94.84	Macao-China	MAC	808	93.63
Switzerland	CHE	5498	94.10	Mexico	MEX	16160	77.64
Czech Rep.	CZE	4036	94.17	Netherlands	NLD	2628	97.08
Germany	DEU	2978	94.51	Norway	NOR	2603	91.95
Denmark	DNK	2733	93.47	New Zealand	NZL	2930	94.42
Spain	ESP	6855	92.05	Poland	POL	2812	93.45
Finland	FIN	3811	95.68	Portugal	PRT	2882	89.84
France	FRA	2755	92.57	Russian Fed.	RUS	3565	85.76
U. Kingdom	GBR	6325	95.53	Slovak Rep.	SVK	4697	93.12
Greece	GRC	2719	84.89	Sweden	SWE	2989	91.94
H.K. - China	HKG	2927	94.24	Thailand	THA	3091	85.55
Hungary	HUN	2787	91.80	Tunisia	TUN	2217	67.72
Indonesia	IDN	5130	69.27	Turkey	TUR	2817	83.29
Ireland	IRL	2565	96.03	Uruguay	URY	2872	70.76
Iceland	ISL	2183	94.14	United States	USA	3529	93.46
Italy	ITA	7306	91.19	Serbia	YUG	2683	87.08
Japan	JPN	3047	93.84	**TOTAL**		**169213**	**88.80**

An important point to be underlined is that in PADIF analyses item parameter estimates from the PISA 2003 international calibration is used (OECD, 2005, p.411). In other words, item calibration is not carried out and the measurement model used in PISA analysis is respected completely.

The analyses

PADIF Analyses have been carried on for five main categorizations. Defining a categorization is defining specifications according to which test items are partitioned. It is one of the most important steps in PADIF, because only a meaningful categorization can contribute to identifying sources of differential group performance. Otherwise, results of PADIF would not make sense. Therefore, using judgments of curriculum or content experts or test developers are of crucial importance in defining a meaningful categorization.

In the present study, the categorizations defined by the PISA-2003 are used (OECD, 2009: 193). However, for two unreleased mathematics items, the document does not contain enough information to decide under which category these items should be placed. So, these two items are excluded from the analyses. In total, 82 mathematics items are studied. Table 7 presents five separate categorizations that are studied in this current research. Abbreviations for category names are given in italics.

Table 7.
The Five Categorizations Studied in the PADIF Analyses

Categorization	Categories	Number of items
Item format	Multiple choice (*Mc*)	27
	Short answer (*Sa*)	41
	Extended response (*Er*)	14
Word count	Short (*Sh*)	21
	Medium (*Md*)	33
	Long (*Ln*)	28
Context	Personal and educational (*Pe*)	38
	Public and scientific (*Ps*)	44
Content area	Space and shape (*Ss*)	19
	Change and relationships (*Cr*)	21
	Uncertainty (*Un*)	20
	Quantity (*Qn*)	22
Competencies	Reproduction (*Rp*)	26
	Connections (*Cn*)	37
	Reflection (*Rf*)	19

Item format. Research shows that language is an important factor in mathematics learning, and the use of language in mathematics items can influence students' performance (O'Leary, 2001; Ellerton & Clements, 1991). For example, the questions that ask students to provide not only a simple answer but also an explanation require an extra effort. It is clear that countries using this type of questions in their instructional practices may have a relative advantage in this type of questions, or vice versa. Thus, it is worth investigating the relative performance of countries in various formats of items. In this categorization, Mc (multiple choice) category consists of items that require students to select one of the response options. The category Sa (short answer) includes the items that require students to construct either a numeric answer, a single word or a short phrase. The items that require an extensive writing, showing a calculation or a justification of the solution are subsumed under the Er (extended response) category.

Word count. This categorization is also related to the language factor. The reading load that is required to understand an item may influence students' performance. OECD (2009: 141) reports the correlation coefficient between the number of words in items and item difficulties as 0.28. In this categorization, there are three categories: *Sh* (short), *Md* (medium) and *Ln* (long). Short items in the category *Sh* are those consisting of 50 words or less. The items comprising 51 to 100 words and more than 100 words are indicated as medium and long, and placed under the categories *Md* and *Ln*, respectively. This classification is based on the English version. Although it is possible that the items would not preserve their length under translation to other languages, it can be assumed that the relative length will be preserved. That is, if item A is shorter than item B in the English version, this ordering will probably be preserved in a translated language. Another advantage of classifying with respect to a single language is to keep the classification the same across languages. For example, if an item contains less than 50 words in English but more than 50 in Turkish, then it belongs to the same category Sh in both languages.

Context. PISA classifies the context in which a mathematics problem is situated with respect to its closeness to the student's life. In this study, the items grouped under the *Pe* (personal and educational) category are those items situated in a context that could actually be experienced by many 15-year-olds. The Items with a problem situation that students might be familiar with through their school curricula are also included in this category. *Ps* (public and scientific) category contains the items situated in a context that can be understood as belonging to the outside world of many 15-year-olds. For example, an item that asks students to interpret and make use of the data on the level of carbon dioxide emissions for several countries belongs to that category.

Content area. In PISA, items are organized based on overarching concepts and relations. The items in the category of *Ss* (space and shape) require an understanding of spatial and geometric phenomena and relationships. Knowledge of mathematical manifestations of change, as well as functional relationships and dependency among variables characterizes the items in the *Cr* (change and relationships) category. The items in the category *Un* (uncertainty) involve probabilistic and statistical phenomena and relationships. Finally, the items that require an understanding of numeric phenomena, quantitative relationships and patterns are included in the category *Qn* (quantity).

Competencies. In the PISA-2003 framework, eight essential competencies are defined as the foundations of mathematical proficiency: mathematical thinking and reasoning, mathematical argumentation, modeling, problem posing and solving, representation, symbols and formalism, communication, and aids and tools (OECD, 2009: 31). To answer the PISA mathematics items correctly, students need many of these competencies at various intensities. With respect to these competency requirements, items are partitioned into three categories. The items in the category *Rp* (reproduction) require competencies at a basic level, such as knowing the facts, recalling mathematical objects and their properties, or performing routine procedures. Requirements of the items in the *Cn* (connections) category are relatively higher. The items in this category are non-routine and require students to interpret and integrate the given information to engage in mathematical decision making. The most demanding items are collected in the category *Rf* (reflection). These items are presented in a relatively unstructured situation. Students have to recognize and extract the mathematics embedded in the situation. They then have to develop strategies to reach a solution. Providing proofs and making generalizations are also requirements of the items in this category.

PADIF analyses are carried out separately for each of these categorizations but, as explained above, prior to the analyses, the categorizations having more than two categories are separated into sub-categorizations so that each have exactly two categories. For example, instead of a single analysis of the categorization ‘Word count’ having the three categories (*Sh*, *Md* and *Ln*) three separate analyses are conducted. In the first analysis, short (*Sh*) items constituted the first category and the rest of the items constituted the second category. Similarly, in the second analysis, the categories are medium (*Md*) items versus the items of other two types, and in the third analysis, categories are long (*Ln*) items versus the others. In this way, PADIF analyses are conducted individually for 14 sub-categorizations each having two categories. In each of the analyses E.P. values for each of the participating countries are calculated.

Results

The comprehensive amount of information provided through PADIF analyses is presented by Table 8 and Figure 1, which should be considered as complements of each other. Full country names corresponding to the acronyms used in Table 8 and Figure 1 can be seen in Table 6.

Table 8.
Countries Ordered with respect to Their Excess Percentages

Item Format			**Word Count**			**Context**		**Area of Content**				**Competency**		
Mc	**Sa**	**Er**	**Sh**	**Md**	**Ln**	**Pe**	**Ps**	**Ss**	**Cr**	**Un**	**Qn**	**Rp**	**Cn**	**Rf**
BRA(16.5)	**ITA(9.8)**	**HKG(9.9)**	**JPN(11.7)**	**YUG(10.0)**	**BRA(8.5)**	**JPN(14.3)**	**ITA(8.9)**	**CZE(10.9)**	**NLD(9.3)**	**IRL(13.3)**	**YUG(13.9)**	**DNK(6.5)**	**USA(7.2)**	**TUN(12.1)**
IDN(8.2)	**SVK(8.9)**	**USA(8.8)**	**KOR(10.2)**	**IDN(7.5)**	**FRA(7.6)**	**YUG(7.9)**	**FRA(8.1)**	**RUS(9.9)**	**LIE(9.2)**	**NOR(12.1)**	**SVK(10.3)**	**YUG(6.2)**	**ISL(6.2)**	**ITA(8.1)**
GRC(7.7)	**RUS(8.8)**	**CAN(8.1)**	**NOR(7.0)**	**NLD(7.4)**	**USA(6.9)**	**MEX(6.1)**	**IRL(7.7)**	**SVK(9.9)**	**RUS(7.1)**	**GBR(9.5)**	**MEX(9.3)**	**PRT(4.9)**	**SWE(4.4)**	**BRA(7.6)**
TUR(7.2)	**AUT(7.1)**	**GBR(7.9)**	**DNK(6.2)**	**THA(6.2)**	**PRT(5.7)**	**SVK(5.9)**	**HUN(6.5)**	**JPN(9.7)**	**FRA(6.8)**	**ISL(8.4)**	**CZE(7.0)**	**NLD(4.4)**	**AUT(3.2)**	**IDN(6.9)**
YUG(5.2)	**CZE(7.0)**	**NLD(7.4)**	**ITA(4.5)**	**POL(5.3)**	**IDN(5.4)**	**DNK(5.5)**	**FIN(5.2)**	**CHE(8.4)**	**USA(6.7)**	**CAN(8.0)**	**DEU(6.6)**	**KOR(3.9)**	**AUS(3.1)**	**TUR(6.0)**
THA(3.6)	**CHE(6.9)**	**MAC(7.1)**	**LVA(4.0)**	**SVK(4.9)**	**URY(5.2)**	**THA(3.7)**	**GBR(2.9)**	**ITA(8.3)**	**TUR(5.0)**	**NLD(7.5)**	**URY(5.5)**	**CZE(3.7)**	**IRL(3.1)**	**MAC(5.9)**
JPN(3.3)	**POL(5.2)**	**FIN(5.3)**	**CZE(3.3)**	LIE(4.9)	**ITA(4.9)**	**IDN(3.6)**	**PRT(2.8)**	**KOR(7.5)**	**GBR(4.5)**	**BRA(6.1)**	**LUX(5.2)**	**POL(3.6)**	**HKG(2.5)**	**KOR(5.4)**
KOR(2.9)	**YUG(4.5)**	**IRL(4.9)**	**SWE(2.8)**	**CZE(4.7)**	**IRL(4.8)**	**RUS(3.5)**	LIE(2.7)	**AUT(7.1)**	**BEL(4.1)**	**HKG(5.8)**	**HUN(4.6)**	**DEU(3.2)**	**CHE(2.4)**	**THA(5.2)**
ISL(2.8)	**LUX(4.4)**	**BEL(3.8)**	USA(1.9)	**RUS(4.6)**	**TUN(4.4)**	**HKG(2.7)**	NZL(2.2)	**LVA(6.8)**	**NZL(3.0)**	**AUS(5.7)**	**AUT(4.5)**	**ESP(3.0)**	NLD(2.0)	LIE(4.6)
USA(2.6)	**FRA(4.0)**	**AUS(3.5)**	MAC(1.9)	**TUN(4.6)**	**TUR(4.1)**	**CZE(2.3)**	URY(2.1)	**TUN(5.5)**	**JPN(2.5)**	**SWE(4.9)**	**ESP(4.3)**	**FRA(2.7)**	**CAN(2.0)**	**IRL(4.3)**
TUN(2.2)	**MEX(3.9)**	**ISL(3.5)**	**SVK(1.9)**	**TUR(3.9)**	**GBR(3.4)**	**USA(2.2)**	TUR(2.1)	**HKG(4.8)**	LVA(2.4)	**IDN(4.5)**	**ITA(3.5)**	NOR(2.3)	IDN(1.6)	**GRC(3.6)**
ESP(2.0)	**DNK(3.7)**	**NOR(3.3)**	HKG(1.4)	**URY(3.9)**	**NZL(3.1)**	BRA(1.8)	GRC(1.9)	MAC(4.5)	HUN(2.3)	**NZL(4.4)**	BRA(2.5)	**SVK(2.2)**	GBR(1.5)	**JPN(3.3)**
URY(1.8)	LIE(3.0)	**SWE(2.4)**	ISL(1.0)	**PRT(3.0)**	**ESP(3.0)**	POL(1.7)	**CAN(1.7)**	**POL(4.1)**	DEU(2.2)	**TUR(4.3)**	**CHE(2.1)**	LIE(2.0)	FIN(1.4)	**URY(3.1)**
NOR(1.6)	**DEU(2.4)**	NZL(2.0)	HUN(1.0)	**IRL(2.7)**	**AUS(2.9)**	TUN(1.2)	**AUS(1.6)**	LIE(3.7)	**AUS(2.2)**	**GRC(3.0)**	RUS(2.0)	URY(1.9)	NZL(1.3)	**MEX(2.9)**
MEX(1.4)	URY(2.3)	JPN(1.5)	CHE(0.6)	HKG(2.1)	**MEX(2.8)**	CHE(0.9)	BEL(1.6)	**THA(2.7)**	PRT(2.0)	**PRT(2.8)**	FIN(1.9)	FIN(1.9)	BEL(1.1)	**BEL(2.7)**
CZE(1.2)	HUN(1.2)	LIE(1.4)	NZL(0.3)	MAC(1.9)	LUX(2.4)	AUT(0.5)	MAC(1.4)	**IDN(2.3)**	KOR(1.7)	USA(2.1)	SWE(1.6)	LUX(0.5)	ESP(0.8)	LUX(2.1)
LUX(1.0)	LVA(1.0)	FRA(0.4)		**BEL(1.9)**	HUN(2.4)	SWE(0.4)	LVA(1.4)	DNK(2.3)	**CAN(1.1)**	THA(2.1)	POL(1.0)	CHE(0.4)	RUS(0.6)	**AUS(2.0)**
HUN(0.8)	BEL(1.0)			**CAN(1.9)**	**CAN(2.1)**	ISL(0.2)	DEU(1.2)	GRC(0.7)	IRL(1.0)	FIN(1.8)	TUN(0.6)	LVA(0.1)	HUN(0.5)	NOR(1.3)
CHE(0.6)	PRT(0.8)			GRC(1.7)	LIE(1.8)	KOR(0.2)	NOR(0.7)	YUG(0.7)	URY(0.1)	DNK(1.6)	BEL(0.6)		MEX(0.4)	PRT(0.8)
LIE(0.4)	TUN(0.3)			CHE(1.5)	GRC(1.4)		ESP(0.5)	DEU(0.2)		POL(1.4)	LIE(0.3)		THA(0.4)	GBR(0.7)
NZL(0.2)	IRL(0.3)			FIN(1.0)	NLD(0.4)		LUX(0.5)	MEX(0.2)		MAC(0.5)	GRC(0.1)		SVK(0.2)	NZL(0.0)
DEU(0.0)				DEU(0.9)	DEU(0.4)		NLD(0.2)				FRA(0.0)		GRC(0.0)	POL(0.0)
				MEX(0.8)	BEL(0.1)									
				JPN(0.7)	AUT(0.0)									
				AUT(0.6)										
				DNK(0.4)										
				ESP(0.1)										

Note. Mc: multiple choice, Sa: short answer, Er: extended response, Sh: short, Md: middle, Ln: long, Pe: personal and education, Ps: public and scientific, Ss: space and shape, Cr: change and relationship, Un: uncertainty, Qu: quantity, Rp: reproduction, Cn: connection, Rf: reflection. The countries where the difference between the observed and expected frequencies of students is statistically significant at the 0.01 level of significance are indicated in bold.

(continued)

Table 8

Countries Ordered with respect to Their Excess Percentages (continued)

Item Format			**Word Count**			**Context**		**Area of Content**				**Competency**		
Ss	**Cr**	**Un**	**Sh**	**Md**	**Ln**	**Pe**	**Ps**	**Ss**	**Cr**	**Un**	**Qn**	**Rp**	**Cn**	**Rf**
SVK(-0.3)	KOR(-0.3)	LVA(-1.1)	YUG(-0.1)	BRA(-0.2)	ISL(-0.3)	NLD(-0.2)	ISL(-0.2)	LUX(-0.2)	ISL(-0.2)	MEX(-0.1)	THA(-0.2)	HUN(-0.2)	DEU(-0.1)	HKG(-0.1)
DNK(-0.3)	NZL(-2.0)	ESP(-1.1)	FIN(-0.7)	KOR(-0.4)	FIN(-1.1)	ESP(-0.5)	KOR(-0.2)	NZL(-0.6)	ESP(-0.6)	ESP(-0.8)	DNK(-0.2)	JPN(-0.5)	TUR(-0.4)	LVA(-0.2)
AUT(-0.4)	MAC(-2.0)	PRT(-1.4)	AUT(-0.7)	LVA(-0.7)	SWE(-1.7)	LUX(-0.5)	SWE(-0.4)	FRA(-0.7)	LUX(-1.4)	LUX(-1.7)	NOR(-0.5)	AUT(-0.6)	YUG(-0.6)	CAN(-0.4)
NLD(-0.6)	**CAN(-2.3)**	**IDN(-1.8)**	DEU(-1.0)	SWE(-0.9)	**CHE(-1.8)**	NOR(-0.7)	AUT(-0.5)	TUR(-0.7)	FIN(-1.9)	**URY(-2.7)**	MAC(-0.6)	RUS(-0.8)	CZE(-0.7)	USA(-0.5)
SWE(-0.9)	**FIN(-2.4)**	GRC(-2.0)	RUS(-1.1)	GBR(-1.0)	**THA(-3.2)**	DEU(-1.2)	CHE(-0.9)	FIN(-0.8)	GRC(-2.4)	**TUN(-3.1)**	IDN(-0.7)	HKG(-1.3)	LUX(-0.9)	RUS(-0.9)
PRT(-1.2)	**AUS(-2.5)**	HUN(-2.2)	POL(-1.2)	ISL(-1.1)	MAC(-3.4)	MAC(-1.4)	TUN(-1.2)	URY(-1.2)	**CZE(-2.9)**	**KOR(-3.3)**	IRL(-0.8)	NZL(-1.4)	FRA(-1.3)	FRA(-1.4)
ITA(-1.6)	**THA(-2.7)**	**KOR(-3.7)**	**AUS(-1.4)**	LUX(-1.3)	**POL(-3.5)**	LVA(-1.4)	POL(-1.7)	AUS(-1.4)	**SVK(-2.9)**	**HUN(-3.8)**	**CAN(-2.0)**	**CAN(-1.8)**	LVA(-1.9)	CZE(-1.7)
AUS(-1.7)	**ESP(-3.0)**	**DNK(-3.9)**	LUX(-1.5)	**AUS(-1.9)**	**RUS(-3.6)**	**AUS(-1.6)**	BRA(-1.8)	BEL(-1.6)	**CHE(-3.0)**	**BEL(-4.3)**	ISL(-2.3)	**GBR(-1.8)**	JPN(-1.9)	HUN(-2.0)
POL(-2.0)	**GBR(-3.3)**	**THA(-4.6)**	GRC(-2.2)	**HUN(-3.4)**	**NOR(-3.7)**	BEL(-1.6)	**USA(-2.2)**	**HUN(-2.5)**	**AUT(-3.0)**	**ITA(-4.4)**	**PRT(-2.5)**	SWE(-1.9)	**ITA(-2.1)**	**FIN(-2.5)**
HKG(-2.4)	**SWE(-3.8)**	**DEU(-5.0)**	**FRA(-2.6)**	**NZL(-3.6)**	**HKG(-4.5)**	**CAN(-1.7)**	**CZE(-2.3)**	**PRT(-3.1)**	**DNK(-3.3)**	**FRA(-6.2)**	**USA(-3.3)**	**MEX(-2.3)**	DNK(-2.4)	**DEU(-2.7)**
MAC(-3.0)	**NOR(-4.2)**	**POL(-5.8)**	**BEL(-3.5)**	**NOR(-3.8)**	**LVA(-4.6)**	GRC(-1.9)	**HKG(-2.7)**	**BRA(-3.1)**	**SWE(-3.4)**	LIE(-6.8)	**LVA(-3.4)**	BRA(-2.3)	**POL(-2.6)**	**SVK(-2.9)**
FIN(-3.5)	**TUR(-4.6)**	**TUN(-6.0)**	**ESP(-3.6)**	**FRA(-5.1)**	**DNK(-5.2)**	URY(-2.1)	**RUS(-3.5)**	**ESP(-3.3)**	**POL(-4.2)**	**JPN(-6.9)**	**HKG(-4.7)**	ISL(-2.6)	**NOR(-2.8)**	**SWE(-3.2)**
IRL(-3.6)	**JPN(-4.8)**	**TUR(-6.1)**	**GBR(-3.8)**	**ITA(-8.1)**	**CZE(-6.1)**	TUR(-2.1)	**IDN(-3.6)**	**SWE(-3.8)**	**THA(-5.1)**	**LVA(-8.4)**	**JPN(-5.1)**	**GRC(-2.8)**	MAC(-2.8)	**CHE(-3.4)**
LVA(-3.7)	**ISL(-6.8)**	**RUS(-6.9)**	**MEX(-4.0)**	**USA(-8.7)**	**SVK(-6.6)**	NZL(-2.2)	**THA(-3.7)**	**USA(-4.6)**	**HKG(-5.3)**	**YUG(-8.4)**	**KOR(-5.5)**	**BEL(-3.1)**	**BRA(-3.5)**	**ISL(-3.7)**
CAN(-4.3)	**NLD(-6.9)**	**LUX(-7.9)**	**THA(-4.9)**		**KOR(-8.4)**	LIE(-2.7)	**DNK(-5.5)**	**NOR(-5.0)**	**NOR(-5.5)**	**CHE(-8.6)**	**GBR(-6.1)**	**AUS(-3.6)**	**PRT(-3.7)**	**AUT(-4.0)**
GBR(-4.7)	**BRA(-7.9)**	**CHE(-8.6)**	**CAN(-5.0)**		**JPN(-8.9)**	**PRT(-2.8)**	**SVK(-5.9)**	**CAN(-6.7)**	**TUN(-5.9)**	**DEU(-9.3)**	**AUS(-6.3)**	MAC(-3.7)	**TUN(-4.2)**	**NLD(-4.8)**
FRA(-4.9)	**USA(-8.3)**	**URY(-8.6)**	**NLD(-6.9)**		**YUG(-9.5)**	**GBR(-2.9)**	**MEX(-6.1)**	**ISL(-6.7)**	**BRA(-6.6)**	**AUT(-10.0)**	**NZL(-8.4)**	**TUR(-3.8)**	**URY(-4.3)**	**ESP(-4.9)**
RUS(-5.5)	**HKG(-8.3)**	**MEX(-9.4)**	**BRA(-7.6)**			**FIN(-5.2)**	**YUG(-7.9)**	**GBR(-7.2)**	**IDN(-7.2)**	**CZE(-14.4)**	**NLD(-8.5)**	**ITA(-4.2)**	LIE(-4.7)	**YUG(-5.0)**
BEL(-5.8)	**GRC(-8.9)**	**AUT(-9.7)**	**TUR(-8.0)**			**HUN(-6.5)**	**JPN(-14.3)**	**NLD(-8.0)**	**YUG(-7.4)**	**SVK(-17.1)**	**TUR(-8.9)**	**THA(-4.4)**	**KOR(-6.7)**	**DNK(-5.0)**
	IDN(-9.2)	**ITA(-11.0)**	LIE(-8.1)			**IRL(-7.7)**		**IRL(-13.2)**	**ITA(-7.5)**	**RUS(-20.2)**		**TUN(-5.1)**		
		CZE(-11.2)	**IRL(-8.2)**			**FRA(-8.1)**			**MAC(-8.1)**			**IDN(-6.0)**		
		SVK(-11.7)	**PRT(-8.8)**			**ITA(-8.9)**			**MEX(-9.9)**			**USA(-7.4)**		
		YUG(-13.6)	**URY(-9.1)**									**IRL(-8.5)**		
		BRA(-15.4)	**TUN(-9.6)**											
			IDN(-14.0)											

Note. Mc: multiple choice, Sa: short answer, Er: extended response, Sh: short, Md: middle, Ln: long, Pe: personal and education, Ps: public and scientific, Ss: space and shape, Cr: change and relationship, Un: uncertainty, Qu: quantity, Rp: reproduction, Cn: connection, Rf: reflection. The countries where the difference between the observed and expected frequencies of students is statistically significant at the 0.01 level of significance are indicated in bold.

Table 8 gives the EP values (in parentheses) in the corresponding category for each of the participating countries. In the table the countries are sorted in descending order of their EP values in each category. In addition, for each category, the countries where the difference between the observed and expected frequencies of students is statistically significant at the 0.01 level of significance are indicated in bold. Notice that in some countries, statistical significance is not reached despite the considerable EP values because of the small sample sizes. A typical example is Liechtenstein (LIE, n=215) where significance was reached for only one of the categorizations (*Cr*).

The EP values can be considered as an indicator of the relative strengths (or weaknesses) of countries regarding the categories. The results show that there are systematic differences among the countries in all five categorizations. It is likely that these differences are due to the cultural differences (e.g., differences in instructional practices, learning experiences, language, educational system etc.) across the countries. However, it is not easy to see the possible patterns in Table 8 among the EP values of the countries. Multidimensional scaling (MDS) may facilitate detecting such patterns by producing a map of the countries with respect to the similarity among their EP values. In Figure 1, such a map is given. The map is produced by the program PERMAP (Heady & Lucas, 1997).

To produce the map, the EP values of countries in each of the categories are defined as the quantified attributes of the countries. However, for the categorization 'context', only the *Pe* category is used; because EP values in the *Ps* category is just the additive inverse of the EP values in the *Pe* category. Thus, in total 14 attributes are defined. PERMAP computes a dissimilarity value for each pair of the countries using these 14 quantified attributes. The simple Euclidean distance in 14 dimensions is used as the dissimilarity measure.

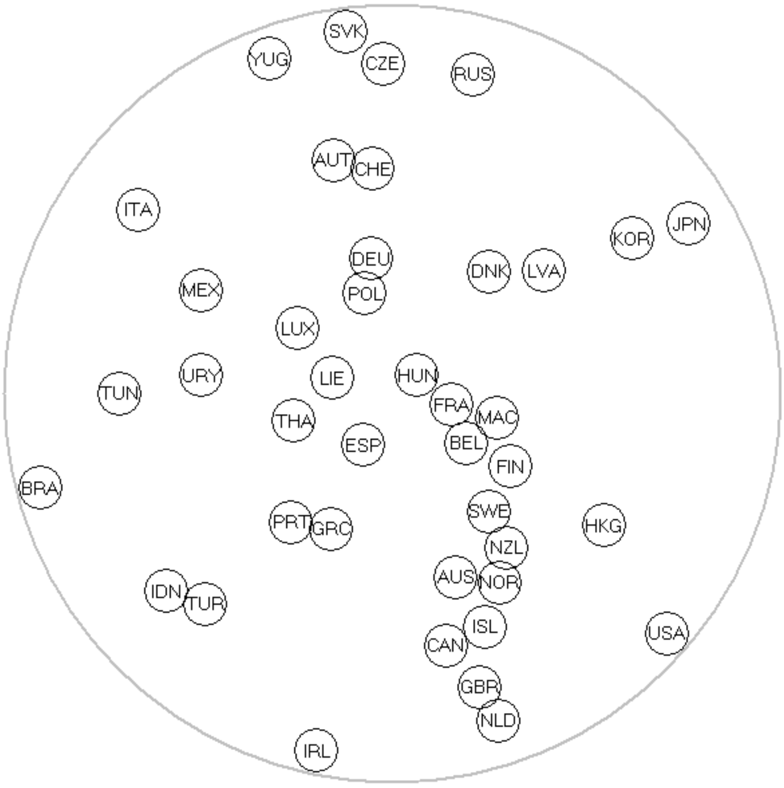

Figure 1. Countries mapped with respect to the similarity of their E.P. values in all 14 categories.

Ratio type MDS algorithm is used in the analysis (Heady & Lucas, 1997). Through this algorithm the program finds a 'map' of the countries in *p*-dimensional space where the distances between the countries are reproduced to be proportional to the corresponding dissimilarity values. In the present study, the analysis is conducted in 3-dimensional space. However, PERMAP can show the results only in plane by showing the 2 dimensions through a method called maximum cross-section. This map is given in Figure 1. The fit measures of PERMAP indicate that the distances in *p*-dimensional space are reasonably reproduced on the plane. The R^2, a coefficient of determination value, is 0.878 and the badness-of-fit value (which is called Stress badness) is 0.015 (Heady & Lucas, 2010).

One can observe that on the map, the countries with similar EP values are placed relatively close to each other and the countries with different EP values are placed relatively far from each other. In addition notice that the countries of similar cultures tend to have similar EP values and therefore tend to be located relatively close to each other in Figure 1.

For example, the Czech Republic (CZE) and the Slovak Republic (SVK) have similar excess percentages in almost all categories. In other words, these two countries have similar strengths or weaknesses on the categorizations used in this study. (CZE and SVK have excess percentages in opposite directions only for *Mc* and *Cn*. But these excess percentages are not statistically significant.) Before the disintegration on 1993, the Czech Republic and the Slovak Republic were combined under the state of Czechoslovakia for almost one century. It is highly possible that these two countries still have similar cultural characteristics, such as similar instructional practices. In line with this, the relative performances of these two countries are also parallel. However, there are some unexpected results as well. For example, it is surprising that Macao-China (MAC) and Hong-China (HKG) are more similar to the Western Countries than they are to their eastern neighbours Korea and Japan. To search for the possible reasons, a follow up in depth analysis in each country can be carried out. This type of research would reveal important cultural characteristics that have an effect on students' performance.

When the results of PADIF are further investigated, parallelism with the findings of earlier studies can be observed. For example, as Wolf (1998) noted, students tend to perform better on the item types they are familiar with. Wolf (1998) also mentions that multiple choice questions are widely used in the United States and that many European countries use constructed response items. One can observe on Table 8 that the relative performances of countries in category *Mc* are in line with this detection. The relation is even more apparent if European countries are limited to the Western European countries.

Similarly, O'Leary (2001) gives the information that examinations in Ireland are dominated by short-answer and essay type questions. He also reports that Irish students performed better on the extended-response items in the Third International Mathematics and Science Study (TIMSS). In line with this finding, notice that Irish students also performed significantly better on the extended-response items in PISA 2003 as well (EP (*Er*) = 4.9%).

Another similarity between PADIF results and findings in the literature is in the categorization "content area". Linking the PISA results to countries' instructional practices is one of the important research areas of the PISA team (OECD, 2009). In one of the analyses to this purpose, they classified the PISA mathematics items under the classical mathematics topics (e.g., Number, Geometry) and they estimated average item difficulties separately in each of the participant countries. Results show that Number items are relatively easy in Yugoslavia, Russia, the Czech Republic, the Slovak Republic, and Austria. In the PISA 2003, Number items are mostly placed under the overarching category of "quantity". PADIF results also indicate that these five countries perform relatively better on the items under the category "quantity".

One final example to support the PADIF results can be the study of Klieme and Bos (2000) (as cited in Klieme & Baumert, 2001). They investigated instructional practices in Japan and Germany through the analyses of TIMSS—Video material. As a result they evaluate Japanese instruction to be better in preparing students to cope with high-level, cognitively demanding, inner-mathematical tasks, and German instruction better in teaching how to cope with standard tasks embedded in application contexts. In the present study the items including standard tasks are partitioned under the "reproduction" (*Rp*) category and the most cognitively demanding items are under the "reflection" (*Rf*) category. When the excess percentages of Japan and Germany on these two categories are observed, one can see that, in line with the evaluation of Klieme and Bos (2000), Germany has performed relatively better in the category reproduction (EP (*Rp*) = 3.2%) while Japan has performed relatively better in the category reflection (EP (*Rf*) = 3.3%).

All these similarities with the findings in the literature can be regarded as evidences for the validity of PADIF results. What's more is that PADIF has a potential to provide very comprehensive amount of information from different perspectives in a single study.

Conclusion

Zumbo (2007) summarizes three major trends in the history of DIF analyses. As a current and future direction, he claims that one of the important uses of DIF is and will be investigating the cognitive and psychosocial processes of item responding across different groups. PADIF might be a promising method serving to this praxis of DIF due to its two main advantages. First, as long as meaningful item categories are defined, PADIF results provide information on the possible sources of DIF. However, this mostly depends on how well the categorizations of items are carried on. In addition to using judgments of experts, some exploratory approaches can also be used to detect item categories. For example some distinct dimensions can be determined through factor analysis, cluster analysis, or multidimensional scaling. Second, PADIF is very flexible. That is, in PADIF the student level analyses can be aggregated at any specific group of individuals. In this current study student level analyses are aggregated at the country level. It is also possible to aggregate the results at other specific groups as well. For example, one can aggregate the results at the student groups of various socioeconomic statuses, or of various teaching practices. Thus, PADIF might produce a broad range of information to shed light to the influence of such contextual variables on students' performance on tests.

Additionally, PADIF produces comprehensive amount of information. Classical dichotomy of reference and focal groups is not an issue in PADIF. So, PADIF can produce information for all available groups of individuals and for all available categorizations of items at a time. One should also notice that PADIF analyses can be carried out to produce more specific information than the one presented in this article. For example, in PADIF analyses with more than two categories, students' performance in all categories can be evaluated at a time. Thus, among the students who perform similarly on a certain category, observing differences in their performance on the other categories may lead important information. Besides, more specific categorizations on certain areas can be defined. For example, all categorizations can be defined considering the language factor. This may contribute to a deeper understanding of the effect of specific factors on the students' responses. In addition, following the international analyses, PADIF can further be conducted in each country with various groups within that country as well. This may provide more specific information on strengths and weaknesses of that country. As a conclusion, the results support PADIF to be considered as a useful approach especially in the context of large scale assessments. It can be used as a routine supplementary analysis to shed light to some country-specific (or any other group-specific) strengths and weaknesses which could not be accounted by the measurement model.

On the other hand, PADIF has some limitations as well. First, the current version of the software used in this study can only be used in tests where the measurement model used is an IRT model in which a total test score is defined. However, research is going on to see how PADIF can be used with measurement models other than the Rasch model like the 2PLM and the 3PLM. Second, evaluation of observed profiles is based on a comparison to the expected profiles. However, if bias is so pervasive in a test that it affects almost all the items, expected profiles might not be a meaningful criteria. Thus, it should be considered that the EP values obtained in PADIF analysis is a valid indicator of strength or weakness only if the measurement model yields valid results. But one should also notice that this is the problem for all DIF techniques that rely on an internal estimate of ability to match examinees.

References

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage Publications.

Ellerton, N.F. & Clements, M. A. (1991). *Mathematics in language: A review of language factors in mathematics learning.* Geelong, Victoria: Deakin University Press.

Heady, R. B., & Lucas, J. L. (1997). Permap: An interactive program for making perceptual maps. *Behavior Research Methods, Instruments, & Computers, 29(3),* 450-455.

Heady, R. B., & Lucas, J. L. (2010). MDS analysis using Permap 11.8. Retrieved from http://www.ucs.louisiana.edu/~rbh8900/

Klieme, E. & Baumert, J. (2001). Identifying national cultures of mathematics education: Analysis of cognitive demands and differential item functioning in TIMSS. *European Journal of Psychology of Education, 16(3),* 385 – 402.

Klieme, E., & Bos, W. (2000). Mathematikleistung und mathematischer Unterricht in Deutschland und Japan:Triangulation qualitativer und quantitativer Analysen am Beispiel der TIMSS-Studie. *Zeitschrift fürErziehungswissenschaft,*3.

O'Leary, M. (2001). Item format as a factor affecting the relative standing of countries in the Third International Mathematics and Science Study (TIMSS). Paper presented at the Annual Meeting of the American Educational Research Association, Seattle, WA.

OECD (2005). *PISA 2003 technical report*. Paris: OECD.

OECD (2009). *Learning mathematics for life: A perspective from PISA*. Paris: OECD.

Oshima, T. C., Raju, N. S., Flowers, C. P., & Slinde, J. A. (1998). Differential bundle functioning using the DFIT framework: Procedures for identifying possible sources of differential functioning. *Applied Measurement in Education, 11*, 353 – 369.

Pae, T. I. & Park, G. P. (2006). Examining the relationship between differential item functioning and differential test functioning. *Language Testing, 23(4)*, 475 – 496.

Roussos, L. & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20*, 355 – 371.

Verhelst, N. D. (2012). Profile analysis: a closer look at the PISA 2000 reading data. *Scandinavian Journal of Educational Research,* 56 (3), 315 – 332.

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

Yıldırım, H. H., & Yıldırım, S. (2011). Correlates of communalities as matching variables in differential item functioning analyses. *H. U. Journal of Education,* 40, 386 – 396.

Zumbo, B. D. (2003). Does item-level DIF manifest itself in scale-level analyses? Implications for translating language tests. *Language Testing, 20*, 136 – 147.

Zumbo, B. D. (2007). Three generations of DIF analyses: considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4,* 223 – 233.