Hacettepe Üniversitesi Eğitim Fakültesi Dergisi (H. U. Journal of Education) 40: 386-396 [2011]

CORRELATES OF COMMUNALITIES AS MATCHING VARIABLES IN DIFFERENTIAL ITEM FUNCTIONING ANALYSES

FARKLI İŞLEYEN MADDE ANALİZLERİNDE ORTAK ETKEN VARYANSIYLA İLİŞKİLİ EŞLEME DEĞİŞKENLERİ

Hüseyin H. YILDIRIM[*], Selda YILDIRIM[**]

ABSTRACT: Multivariate matching in Differential Item Functioning (DIF) analyses may contribute to understand the sources of DIF. In this context, detecting appropriate additional matching variables is a crucial issue. This present article argues that the variables which are correlated with communalities in item difficulties can be used as an additional matching variable in DIF analyses. To examine this claim, mathematics data from the Program for International Student Assessment (PISA) was analyzed. Out of school time students spent for learning was detected as an additional matching variable. Multivariate DIF results showed that this variable might be related to the source of DIF in some items.

Keywords: multivariate matching, differential item functioning, logistic regression, PISA

ÖZET: Farklı işleyen madde analizlerinde, birden fazla eşleme değişkeni kullanmak, sorunun farklı işleme sebeplerinin anlaşılmasına katkı sağlayabilir. Bu bağlamda, kullanılabilecek ek eşleme değişkenlerinin tespit edilmesi önemli bir konudur. Bu çalışma, madde güçlüklerindeki ortak varyansla ilişkili değişkenlerin, farklı işleyen madde analizlerinde ek eşleme değişkeni olarak kullanılabilirliğini incelemektedir. Bu amaçla, Uluslararası Öğrenci Başarısını Değerlendirme Programı (PISA) verileri incelenmiştir. Öğrencilerin okul dışında öğrenme için harcadıkları zaman, kullanılabilir bir ek eşleme değişkeni olarak tespit edilmiştir. Çok değişkenli eşleme yöntemiyle, bu değişkenin bazı sorulardaki farklı işlemenin sebebi olabileceği görülmüştür.

Anahtar sözcükler: çok değişkenli eşleme, farklı işleyen madde analizi, lojistik regresyon, PISA

1. INTRODUCTION

In cross-cultural studies, such as Program for International Student Assessment (PISA), many pieces of evidence are needed to support the claim that test scores are unbiased. Results from Differential Item Functioning (DIF) analyses are among such evidence. DIF analyses investigate whether examinees at the same level on the construct being measured by the test but from different countries (herein called same-abilities) have the same probability, on average, of correctly responding to an item (Camilli 2006; van de Vijver & Leung 1997). If not, the item is labeled as "item showing DIF", or shortly, "DIF-item". DIF in an item may be interpreted that in addition to the construct being measured by the test, there is at least one additional country specific factor affecting performance of the same-abilities on that item. However, identifying these specific factors still remains to be a challenging issue in DIF analyses. This current study provides a procedure which can be used for this purpose.

In DIF analyses, the same-abilities can be specified through two methods: Univariate matching and multivariate matching. In the univariate matching, usually, test scores or latent ability scores of individuals are used as a matching variable. That is, the individuals from different countries but having similar total test scores are specified as the same-abilities. It is also possible to specify the same-abilities with respect to more than one matching variable. This approach is known as the multivariate matching, and it may have two advantages as compared to the univariate matching. First, it improves matching. Flagging an item as showing DIF through the multivariate matching, provides stronger evidence as compared to the univariate matching to claim that the students' country membership has an effect on their item responses (Wu & Ercikan 2007).

Second, it may help discern possible sources of DIF. If an item shows DIF in the univariate but not in the multivariate matching analysis, it can be argued that the dimension represented by the

[*] Yrd. Doç. Dr., Abant İzzet Baysal Üniversitesi, e-posta: yildirim.huseyin@ibu.edu.tr
[**] Yrd. Doç. Dr., Abant İzzet Baysal Üniversitesi, e-posta: cet_s@ibu.edu.tr

additional matching variable may be related to the sources of DIF in that item (Clauser, Nungester, & Swaminathan 1996; Zwick & Ercikan 1989). For example, if DIF in an item of a mathematics test disappears after using, say, reading scores of individuals as a second matching criterion, this may provide an indication that reading performance is, somehow, associated with the differential performance of the same-abilities on that item.

Despite its significance, detection of the additional variables to be used as a second matching variable in the multivariate DIF analyses does not have convincing routine procedures yet. Among the frequently used methods are factor analysis, cluster analysis and multidimensional scaling. In addition, a detailed analysis of the content areas in test specifications, or cognitive task analyses are also used in revealing additional dimensions (Gierl 2005; Robin, Sireci & Hambleton 2003). However, all these methods are useful to an extent that the dimensions identified by these procedures are interpretable, which is an issue to be decided through a subjective process, and these subjective interpretations are usually unreliable among judges or inconsistent with DIF statistics (Gierl 2005; Gierl & Bolt 2003; Kupermintz, Ennis, Hamilton, Talbert & Snow 1995; Roussos & Stout 1996).

This current paper presents a procedure to detect a second matching variable, in a way to control possible unreliabilities due to subjective interpretations. The procedure is inspired from the study by Grisay and Monseur (2007). In their study, the authors used complement of communalities from Principal Component Analysis (PCA) as indicators of the amount of DIF. This current study investigated whether variables, which were correlated with communalities in item difficulties might serve as additional matching variables in DIF analyses. There are various methods to be used in DIF analyses, such as Mantel-Haenszel method, Item Response Theory (IRT) based methods or Logistic Regression (LR) method (Camilli & Sheapard 1994). The advantage of LR is that it has a relatively more flexible algorithm that allows using more than one matching variable. Therefore, DIF analyses in this study were conducted through the LR method.

2. METHODOLOGY

2.1. The Data

The data used in DIF analyses consisted of individuals' responses to mathematics tests of the PISA 2003 and PISA 2006. In addition, to specify the country level variables, data from the PISA 2006 student questionnaire was used.

PISA is an ongoing internationally standardized assessment which is carried out by the Organization for Economic Co-operation and Development (OECD). Starting from 2000, it is administered in every three years to assess how well 15-year-old-students are prepared to meet the challenges of knowledge societies (OECD 2005). Forty-one countries participated in the PISA 2003 and 57 countries in the PISA 2006. Forty countries that participated in the PISA 2003 also participated in the PISA 2006. In this study, the data from these 40 countries was used. The tests in PISA were typically administered to between 4,500 and 10,000 students in each country. In most countries, students were selected through the two-stage stratified sampling. That is, sampling units were individual schools, at the first-stage, and students within sampled schools, at the second-stage (OECD 2005).

PISA assessed how far students have acquired some essential, information society-required knowledge and skills in the domains of reading, mathematical and scientific literacy (OECD 2005). PISA also collected information on student and school characteristics through questionnaires.

PISA used a rotated test design in producing booklets to assess achievement. One of these booklets was randomly assigned to each of the sampled individuals. Both in the PISA 2003 and PISA 2006, 13 booklets were produced. In both administrations, each test item appeared in four of the booklets. This linked design enabled estimation of item difficulties on a common scale (OECD 2005).

In the PISA 2003, 84 mathematics items were used. Forty eight of these items were also used to build mathematics tests of the PISA 2006. All 13 booklets of the PISA 2003 included mathematics items. The number of mathematics items in these booklets ranged from 12 to 37. In the PISA 2006, 10

of 13 booklets included mathematics items. The number of mathematics items in these booklets ranged from 12 to 24.

PISA mathematics items varied in format; in addition to multiple choice items, PISA also included free-response items, which required students to construct their own responses. In this study, for free response items scored on 0 to 2 scale (0 indicating a wrong answer, 1 indicating a partially correct answer, and 2 indicating a totally correct answer), the scores were rescaled to a scale of 0 to 1 prior to analyses so that the interpretation of the statistics would be the same for all item types. In this process, both the partially correct answers and the totally correct answers were treated as correct answers and coded as 1.

This study also used data from the student questionnaire of the PISA 2006 to specify country-level variables. This questionnaire includes questions to collect information on the educational background characteristics of the countries. Three questions in this questionnaire were detected due to their possible effect on mathematics achievement. These three questions asked students to specify the typical amount of time they spent per week for, a) regular lessons in school, b) out-of school-time lessons, and c) self-study or homework. For each question, students were to select one of five choices indicating the amount of time from 'no time', to 'six or more hours a week'. These choices were coded from 1 to 5, respectively.

2.2. The Procedure

The analyses conducted in this study consisted of three phases. These three phases are briefly described here and details of the analyses conducted at each phase are given in the following sections.

At the first phase, item difficulties were estimated for each country separately. To this purpose, for each country, their data from the PISA 2003 and PISA 2006 were merged. This procedure increased the number of individuals who responded to the items and thus led more precise item difficulty estimations.

The second phase included estimation of the communalities in the item difficulties at the country level. Then, correlations between these communalities and the country-level variables obtained from the student questionnaire were investigated to detect possible candidates for the second matching variable.

At the last phase of the study, two DIF analyses were conducted through the use of LR method in the PISA 2006 mathematics data. In the univariate matching case, the total test score (number of correct responses) of students was used as a matching variable. In the multivariate matching case, the country-level variables detected at the second phase were used in addition to the total test score. The results from two DIF analyses were compared.

2.3. The Analyses

2.3.1. Item Difficulty Estimations

One Parameter Logistic Model (OPLM), which is an IRT model, was used to estimate the item difficulty parameters (Verhelst, Glass, & Verstralen 1991). Information on this model and the software is as follows.

OPLM is a hybrid model that combines the appealing theoretical advantages of the Rasch Model and the flexibility of Two Parameter Logistic Model (2PLM) in handling unequal discriminations of the items (Verhelst & Glass 1995). OPLM is formally identical to the 2PLM whose item response function is;

$$f_i(\theta) = \frac{\exp[\alpha_i(\theta - \beta_i)]}{1 + \exp[\alpha_i(\theta - \beta_i)]}, \text{ for } i = 1,.....,k, \qquad (1)$$

where $\alpha_i > 0$ is the discrimination and β_i is the difficulty parameter of an item *i*.

However, the discrimination indices α_i in OPLM are not unknown parameters, but hypothesized integer constants. Using these constants as weights makes the weighted sum score of individuals a

sufficient statistic for abilities, θ, as in the Rasch Model. This enables obtaining the Conditional Maximum Likelihood (CML) estimates of item difficulties, which otherwise would only be used in the Rasch Model (Verhelst & Glass 1995). Using CML has the following advantages: a) Item difficulties can be estimated without any assumption on the ability distribution of individuals, b) the item calibration in incomplete designs, such as in PISA, is possible, and c) test statistics at the item level are available (Molenaar 1995).

OPLM software has a built-in function to obtain initial values for discrimination indices (Verhelst, Verstralen & Egen 1991). Using these discrimination values as fixed indices, CML estimates of the item difficulties can be obtained. OPLM software also provides statistical tests on suitability of the discrimination indices so that one can adjust these values to get satisfactory results.

In CML estimation, the OPLM software produces M_i and S_i tests. The subscript indicates that the tests are item oriented. These tests have power against the misspecification of discrimination indices. The M_i test also provides a suggestion on (upward or downward) the adaptation of discrimination indices to get a better fit. In addition, the OPLM software produces R statistic, which is a global test for the model fit. All these tests belong to the family of generalized Pearson tests introduced by Verhelst and Glass (1995).

In this study, the test statistics of each of the 84 items on the suitability of fixed discrimination indices were investigated, and adjusted when required. Finally, each of the item characteristic curves, which display the difference between observed and expected probability of the correct response with respect to ability levels, were investigated to check whether there were severe problematic items. These analyses were conducted separately for each of the 40 countries.

2.3.2. Communalities and Correlates

The country communalities were estimated through PCA. Mathematics items were specified as the observations, and the countries were specified as the variables. The values of the variables were the IRT item difficulties, which were estimated at the first phase separately for each country. The communalities of each country can be used as an indicator of the proportion of variance in the item difficulties in that country that can be accounted to the common construct being measured by PISA (Grisay, de Jong, Gebhardt, Berezner, & Halleux 2006). The underlying rationale is given below.

In unidimensional IRT framework, item difficulties specify positions of the items on the continuum of the latent construct, which is "mathematical literacy" in the context of PISA (OECD 2005). Item difficulties are estimated from the response data of individuals. It is assumed that the ability of an individual, i.e. the position of the individual on the latent construct, is the only factor affecting the response of the individual on an item (Hambleton, Swaminatthan & Rogers 1991; OECD 2005). Besides, in IRT framework, the item parameters estimated from different samples are invariant up to a linear transformation, provided that IRT model fits the data (Hambleton, Swaminathan & Rogers 1991).

Thus, if the assumptions specified above holds perfectly, in each country, all the variance in item difficulties should be accounted by the performance of individuals, producing communality values of 1 (corresponding to 100%) for each country. However, the existence of some country-specific factors, which affect performance of individuals, may result in a loss in the value of communalities (Grisay & Monseur 2007).

In the same manner, an association between some country-level variables and country communalities can be regarded as evidence that these country-level variables have a direct or an indirect effect on the performance of students. Thus, these country level variables can be used as additional matching variables in DIF analyses.

In this study, the correlations between country communalities and the average time that the students in a country spent per week for: a) regular lessons in school, b) out-of school-time lessons, and c) self-study or homework were investigated. Variable(s) with significant correlation with communality in item difficulties was further investigated in DIF analyses as a second matching variable.

2.3.3. DIF Analyses

The Logistic Regression (LR) method was used in DIF analysis. LR provides an advantage of using more than one variable in matching individuals from different groups. Swaminathan and Rogers (1990) presented the LR model which gives the probability of a correct response to an item as:

$$P(u = 1|\theta, g) = \frac{e^{\tau_0 + \tau_1\theta + \tau_2 g + \tau_3(\theta g)}}{1 + e^{\tau_0 + \tau_1\theta + \tau_2 g + \tau_3(\theta g)}} \qquad (2)$$

θ in the model is the matching variable which is usually called observed ability. The group membership is specified with a categorical variable, g. θg in the model represents the ability-group interaction. The parameter τ_0 is the intercept and τ_1, τ_2 and τ_3 are the weights. In this current study, θ is the number of correct responses of an individual.

In LR a significant non-zero value of τ_2 is judged as evidence of uniform DIF. Similarly a significant non-zero value of τ_3 is judged as evidence of non-uniform DIF. The difference between the -2 log likelihood of the compact model that fixes the value of one or both of these weights to zero and the augmented model that releases at least one of these limitations is used as a chi-square test statistics with degree of freedom (*df*) equal to the difference between the number of free weights in the models (Camilli & Sheapard 1994).

A second matching variable θ' can be included in the model (2). In this case, the model is to include four interaction terms: namely, the interaction between two matching variables, two interactions between matching and grouping variables separately, and the combined interaction between two matching and grouping variables.

To quantify the magnitude of DIF, effect size classification guidelines are proposed for each step of LR DIF procedure in the literature (Jodoin & Gierl 2001; Zumbo 1999). In this study, Jodoin and Gierl's (2001) effect size measure, R^2, was used. According to this measure, DIF statistics with a R^2 value smaller than .035 is considered as negligible and denoted as A-DIF. Moderate DIF is denoted as B-DIF and corresponds to R^2 value between .035 and .070. Finally, R^2 value greater than .070 specifies high DIF denoted as C-DIF.

3. RESULTS

The item difficulties estimated through the use of OPLM were in line with the estimates reported in the PISA 2003 technical report, producing a significant Pearson correlation ($r = .947$, $p < .01$). In PCA, there were 40 variables (countries) and 84 observations (items). Values of variables were the item difficulties estimated in the first phase. With respect to Kaiser's (1960) criteria, PCA produced two components whose eigenvalues were greater than 1 (34.133 and 1.280). These two components accounted for 85.33% and 3.20% of variance, respectively. However, with respect to Stevens' (2002) criterion, which considers the significance of factor loadings, none of the loadings on the second factor were significant. On the other hand, all the loadings on the first factor, ranged from .976 to .747, were significant. Thus, communalities were calculated only with respect to the first factor. The communalities are given in Figure 1.

The thin lines on each of the horizontal bars in Figure 1 are the 95% confidence intervals. The confidence intervals were based on the bootstrap methodology as summarized by Timmerman, Kiers and Smilde (2007). Hundred bootstrap samples of size eighty-four were randomly drawn with replacement from the original data of the item difficulties obtained from 40 countries. The standard deviation from the bootstrap distribution (i.e., distribution of communality estimations) of each country was considered as the bootstrap standard error for that country. The confidence intervals were calculated based on the normality assumption.

With respect to the Figure 1, the communalities obtained from most of the western countries were high in magnitude. On the other hand, most of the countries took the versions developed in non-Indo-European languages were placed at near the bottom of the graph. However, Mexico and Brazil were also placed among the low-communality countries, although, they took the versions in Indo-European languages. When performance of countries on mathematics items was considered, most of

the countries placed by the bottom of the graph were either low or high performing countries in the mathematics domain of the PISA 2006. Korea and Japan located among the best performing countries whereas Tunisia, Indonesia, Mexico and Turkey were among the poor performing countries.

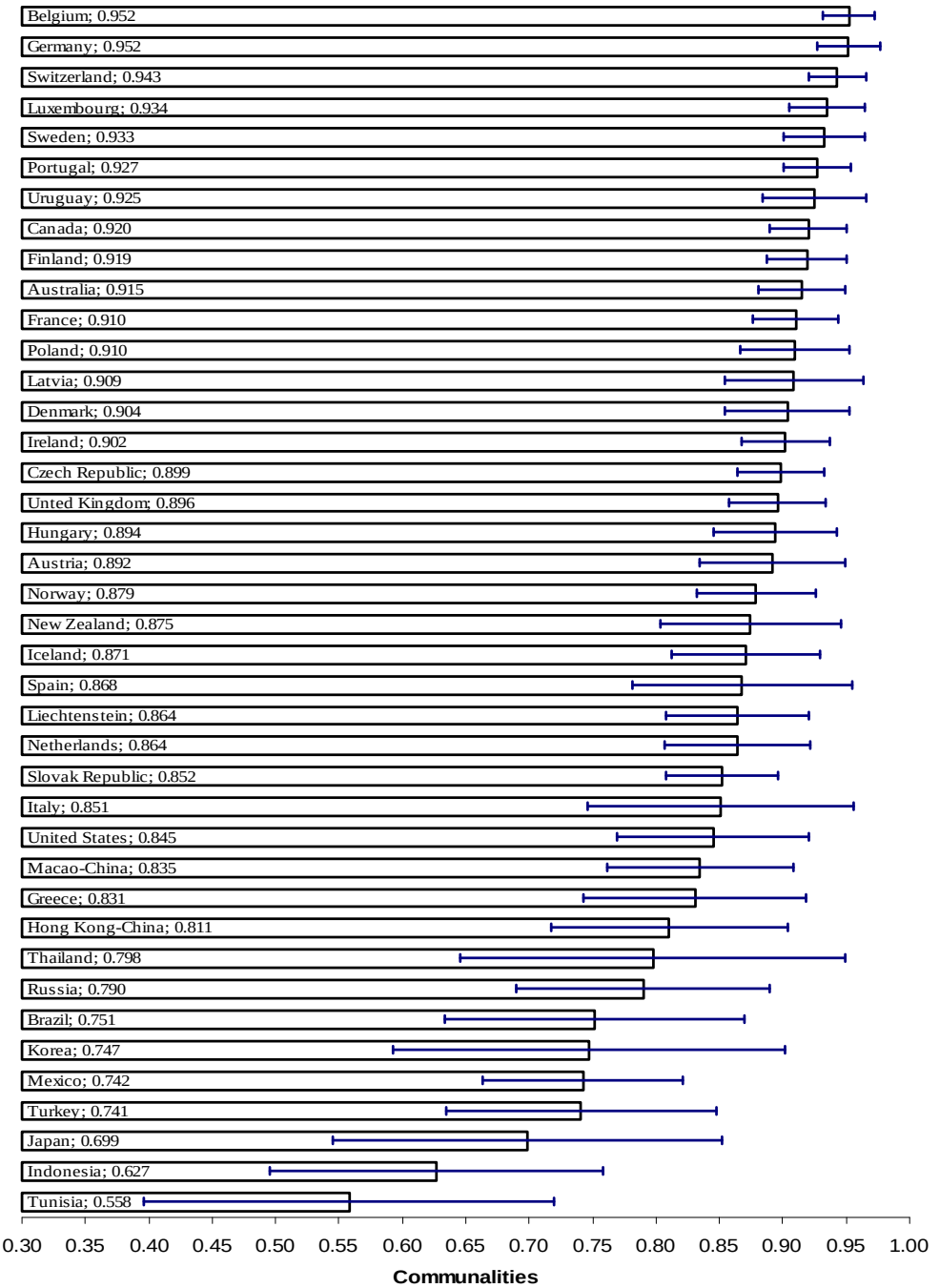

Figure 1. Communalities in Item Difficulties and Confidence Intervals

The correlations between the indicator of communality and the selected background characteristics of 40 countries are given in Table 1. The negative and statistically significant correlation between the communality and out-of school-time confirmed that the countries which had low communality in item difficulties tend to have more out-of school-time lessons than the countries having high communality in the item difficulties. Kolmogorov-Smirnov (K-S) tests showed that the normality assumption of the Pearson correlation was not violated for both variables (Communality, *p*

= .173, K-S z = 1.106; OSL, p = .289, K-S z = .983). In addition, the test of linearity showed that deviation from linearity between the two variables was not statistically significant (p = .444, F = 1.178, df = 31).

Table 1. Correlations between Communalities and Time Spent on Learning

N= 40 countries	RL	OSL	SSH
Communality	-.033	-.632**	-.399*

Note. RL: Regular Lessons; OSL: Out-of School-time Lessons; SSH: Self-Study or Homework
** $p < .01$, * $p < .05$

The Pearson correlation in Table 1 can be interpreted that the out-of school-time lessons variable may affect the performance of individuals, thus, affecting item difficulties in some countries. To investigate this claim, DIF analyses between two selected countries were conducted.

Finland and Korea were selected for DIF analyses. The rationality of selecting these countries was as follows. With respect to the PISA 2006 mathematics scale, these two countries were among the best performing countries with almost similar means 548 and 547, respectively. Thus, selecting these countries, the possible effect of the difference between performances of countries on DIF results would be controlled to an extent. On the other hand, these countries were on opposite edges with respect to their average scores on out-of school-time lessons variable. This provided an information rich case to see the possible effect of this variable on differential performances of individuals.

The average scores of Finland and Korea on the out-of school-time lessons variable were 1.3 and 2.5, respectively. Scores of the 40 countries on this variable ranged from 1.3 to 2.6, with the mean 1.77 and the standard deviation .31.

DIF analyses for Finland and Korea on the PISA 2006 mathematics data (47 items) were conducted on six booklets, separately. As a consequence of PISA test design, each of the 47 items was appeared in three of the six booklets. The behavior of an item in each of the three booklets it appeared was investigated. There were approximately 400 Korean students and 350 Finnish students responding to each booklet.

Results of DIF analyses from univariate (only total score of individuals as the matching variable) and multivariate (both total test score and time individuals spent out-of school for lessons are matching variables) matching analyses are given in Table 2.

Table 2. PISA 2006 Mathematics Items Flagged as DIF

		Matching Variables[a]					
Items	Booklets	TS	TS&OSL	TS	TS&OSL	TS	TS&OSL
M12	3,7,10	C	C	C	B	C	C
M15	4,10,13	C	C	C	C	C	C
M20	3,8,13	B	B	C	B	B	A
M21	3,8,13	C	C	C	C	C	C
M25[b]	**4,10,13**	**C**	**A**	**C**	**B**	**B**	**A**
M26	4,10,13	B	B	C	B	B	A
M27	4,7,8	C	C	C	C	C	C
M29	**4,7,8**	**C**	**B**	**B**	**A**	**C**	**B**
M40	4,10,13	C	B	B	B	C	C
M47	4,10,13	C	C	C	C	C	C

Note. TS: DIF results with Total Score as single matching variable; TS&OSL: DIF results with Total Score and Out-of School-time Lessons as matching variables; C: Large-DIF; B: Moderate-DIF; A: Negligible-DIF

a. Consecutive pairs of columns TS and TS&OSL, present results from three booklets specified in Booklets column, respectively.

b. Bold rows indicate items whose DIF magnitude reduced in all three booklets in two matching case.

In the univariate DIF analyses, the items flagged as showing DIF in one or two of the booklets but not in the others were ignored due to the inconsistency. As a result, ten of the 47 items showed either moderate or high DIF in all three of the booklets they appeared. When the magnitudes of these ten flagged items were considered, five of the flagged items were at C-DIF level in all three of the booklets they appeared. Three of the flagged items were at C-DIF level in two of the booklets and two of the flagged items were at C-DIF level in one of the booklets.

Finally, multivariate DIF analyses were conducted. The magnitude of DIF in two of the ten items was decreased in all three booklets they appeared when the score of individuals on out-of school-time variable was used as a second matching variable in addition to the total test score.

4. CONCLUSION AND SUGGESTIONS

The results presented above can be further evaluated as follows: First, this study used a different model and software to estimate item difficulties than that was originally used to scale PISA data. However, the high correlation between OPLM item difficulty estimates and the estimates published by PISA can be considered as evidence that OPLM also holds to PISA data. Thus, OPLM estimates can be considered free from the threat of the model data misfit.

The second issue is on seeking evidence to support the claim that communalities represent the strength of unidimensionality. In other words, evidence is required to claim that the communality values represent the strength of the relation between items and the construct being measured by PISA. This evidence may be found in the technical report of OECD (2005). This report presents the results of the investigation of the items in the national versions of the test material and gives the proportion of weak items in each country. When countries were compared with respect to the proportion of the weak items published in the report and the communality values estimated in this study, it was detected that the countries with a low proportion of weak items tended to have high communalities (Spearman's rank correlation was .82; $p < .01$). Therefore, the complement of the country communality can be considered as a variance in that country that cannot be explained by the construct measured by the test. As a consequence, this complement can also be considered as an indicator of the total amount of DIF in that country (Grisay & Monseur 2007).

Finally, investigating the correlates, time spent for out-of-school lessons and time spent for self-study or homework had negative correlations with the communality values. This can be interpreted as, the countries where unidimensionality was satisfied spent lower amount of time in out of school lessons. This finding does not imply causality, but it may be regarded as evidence that time spent on learning out-of-school is, somehow, related to factors causing some items to function differentially across countries. To check this argument, the time students spent for out of school lessons was used as an additional matching-variable in DIF analyses. This reduced the DIF magnitude significantly in two mathematics items. So, the time students spent for out of school lessons may be regarded as a source of DIF between Korea and Finland. Unfortunately, as these two items are not released, it was not possible to investigate the content of these items.

It is worth specifying that Wu and Ercikan (2007) also had a similar result in their study. They investigated the Third International Mathematics and Science Study (TIMSS) 1999 and detected extra school hours as a possible source of DIF between Taiwan and United States.

Depending on the empirical evidence provided above, country communalities seem to capture differences among countries in the extent to which their data departs from unidimensionality. In addition, investigating the correlates of communalities may shed light on possible causes of DIF in some items. In other words, these correlates can be used as additional matching variables in DIF analyses to determine DIF items. The decrease in the amount of DIF may be regarded as evidence of the effect of the additional matching variable on the performance of individuals. Then, further qualitative investigations also may be conducted to disentangle the possible reasons.

Considering PISA mathematics items, approximately 85% of the total variance in item difficulties across the countries was accounted by a common factor. This indicated a substantial comparability of the item difficulties across the countries. However, investigating the communality values of the countries, it is clear that some country-specific factors also had an effect. The amount of

country-specific factors was bigger for the countries that were relatively different on cultural, linguistic, economic or educational characteristics. There is a considerable amount of research in the DIF literature on the effect of such cultural differences (Allalouf, Hambleton, & Sireci 1999; Ercikan, Gierl, McCreith, Puhan, & Koh 2004; van de Vijver & Tanzer 1998).

Finally, regarding the limitations of the method introduced and recommendations for further studies, the following should be specified. The method introduced in the study to identify the second matching variable can be used only when the test includes a reasonable number of items and administered in a reasonable number of countries or groups. Otherwise the number of items would not be enough to conduct PCA, and the number of countries would not be enough to calculate correlations.

There are some theoretical concerns on the estimation of item difficulties. For example, the fluctuation in the model-data fit rate among countries may have an effect of communality estimations or different estimation errors due to different sample sizes of countries may affect estimations. Further studies, such as the simulations controlling these possible fluctuations may clarify some possible defects of the method detailed in the study.

In conclusion, the method introduced in the study seems to produce promising results in identifying some country-level variables that might be associated with the performance of individuals on some items. Thus, the variables correlated with communalities in item difficulties may be good candidates for the additional matching variables in DIF analyses.

REFERENCES

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

Camilli, G. (2006). Test Fairness. In R.L. Brennan (Ed.), *Educational Measurement*. (4th edition, pp. 221-256).Westport, CT: Praeger.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items* (Vol. 4). Thousand Oaks, CA: Sage.

Clauser, B. E., Nungester, R. J., & Swaminathan, H. (1996). Improving the matching for DIF analysis by conditioning on both test score and an educational background variable. *Journal of Educational Measurement, 33(4)*, 453-464.

Ercikan, K., Gierl, M. J., McCreith, T., Puhan, G., & Koh, K. (2004). Comparability of bilingual versions of assessments: Sources of incomparability of English and French versions of Canada's national achievement tests. *Applied Measurement in Education, 17*, 301–321.

Gierl, M. J. (2005). Using dimensionality based DIF analysis to identify and interpret constructs that elicit group differences. *Educational Measurement Issues and Practice, 24 (1)*, 3-13.

Gierl, M. J., & Bolt, D. (2003 April). *Implications of the multidimensionality-based DIF analysis framework for selecting a matching and studied subtest*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Grisay, A., & Monseur, C. (2007). Measuring the equivalence of item difficulty in the various versions of an international test. *Studies in Educational Evaluation, 33*, 69-86.

Grisay, A., de Jong, J. H. L., Gebhardt, E., Berezner, A., & Halleux, B. (2006 July). *Translation equivalence across PISA countries.* Paper presented at the 5th Conference of the International Test Commission, Brussels, Belgium.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. California: Sage.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education, 14*, 329-349.

Kupermintz, H., Ennis, M., Hamilton, L., Talbert, J., & Snow, R. (1995). Enhancing the validity and usefulness of large-scale educational assessments: NELS 88 mathematics achievement. *American Educational Research Journal, 32*, 524-554.

Molenaar, I. W. (1995). Estimation of item parameters. In G.H. Fischer & I.W. Molenaar (Eds.), *Rash models: foundations, recent developments, and applications.* (pp.39-51). New York: Springer.

OECD (2003). *The PISA 2003 assessment framework: mathematics, reading, science and problem solving knowledge and skills*. Paris: OECD.

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

Robin, F., Sireci, S. G., & Hambleton, R. K. (2003). Evaluating the equivalence of different language versions of a credentialing exam, *International Journal of Testing*, 3, 1–20.

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

Stevens, J. (2002). *Applied multivariate statistics for the social sciences*. London: Lawrence Erlbaum.

Swaminathan, H., & Rogers, J. H. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27(4)*, 361-370.

Timmerman, M. E., Kiers, H. A. L., & Smilde, A. K. (2007). Estimating confidence intervals in principal component analysis: A comparison between the bootstrap and asymptotic results. *British Journal of Mathematical and Statistical Psychology, 60*, 295-314.

van de Vijver, F. J. R., & Leung, K. (1997). Methods and data analysis for cross-cultural research. Newbury Park, CA: Sage.

van der Vijver, F., & Tanzer, N. K. (1998). Bias and equivalence in cross-cultural assessment. *European Review of Applied Psychology, 47*, 263–279.

Verhelst, N. D., & Glass, C. A. W. (1995). The one parameter logistic model. In G.H. Fischer & I.W. Molenaar (Eds.), *Rash models: foundations, recent developments, and applications*. (pp.215-237). New York: Springer.

Verhelst, N. D., Glass, C. A. W., & Verstralen H. H. F. M. (1991). *OPLM: a one parameter logistic model for dichotomous and polytomous data*. Measurement and Research Department Reports. Arnhem: Cito.

Verhelst, N. D., Verstralen, H. H. F. M., & Eggen, Th. J. H. M. (1991). *Finding starting values for the item parameters and suitable discrimination indices in the one-parameter logistic model*. Measurement and Research Department Reports. Arnhem: Cito.

Wu, A. D., & Ercikan, K. (2007). Using multiple-variable matching to identify cultural sources of differential item functioning. *International Journal of Testing, 6(3)*, 287-300.

Zwick, R., & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. *Journal of Educational Measurement, 26 (1)*, 22-66.

Genişletilmiş Özet

Uluslararası Öğrenci Başarısını Değerlendirme Programı (PISA) gibi, farklı dil veya kültüre sahip ülkelerde uygulanan testlerden elde edilen sonuçların karşılaştırılabilirliğine dair çalışmalar, son yıllarda, hatırı sayılır bir orana ulaşmıştır. Bu çalışmalarda, test sonuçlarının kültürlerarası denkliğiyle ilgili kanıt üreten, Farklı İşleyen Madde (FİM) analizleri sıkça kullanılmaktadır.

FİM analizleri, temelde, testte ölçülen yeterlik düzeyi denk olan, fakat testin uygulandığı farklı gruplarda yer alan öğrencilerin, sorulardaki performanslarının karşılaştırılmasına dayanır. Yeterlik düzeyi, genellikle, öğrencilerin test puanlarına göre belirlenir. Sonuçta, farklı gruplarda yer alan, ancak test puanları birbirine yakın öğrenciler eşlenmekte ve eşlenmiş öğrencilerin, belirli bir sorudaki performanslarının benzer olup olmadığı incelenmektedir. Bu bağlamda, test puanına, eşleme değişkeni denmektedir. Eşlenmiş öğrenci grupları arasında, belirli bir sorudaki, istatistiksel olarak anlamlı performans farklılığı, söz konusu sorunun, gruplar arasında farklı işlediğine bir kanıt sayılmaktadır. Eşlenmiş grupların bir sorudaki performans göstergesi, kullanılan FİM analizine göre, gruplarda sorunun doğru cevaplanma oranı veya sorunun grup yeterlik düzeyine denk gelen doğru cevaplanma ihtimali olarak alınır (Camilli & Sheapard 1994).

Denk yeterlik gruplarını belirlemek üzere yapılan eşleştirme, birden fazla eşleme değişkenine göre de yapılabilir. Bu yöntemin, tek eşleme değişkeni kullanılan FİM analizlerine kıyasla, iki önemli üstünlüğü vardır: 1) Yeterlik düzeyleri iki değişkene göre belirlendiğinden, grupların yeterlikleri birbirine daha benzer olacaktır. 2) Bir soruda, tek eşleme değişkeni kullanıldığında tespit edilen farklı işleyiş, iki eşleme değişkeni kullanılan FİM analizinde görülmüyorsa, ikinci eşleme değişkeninin gruplar arasındaki performans farklılığıyla ilişkili olduğu iddia edilebilir (Wu & Ercikan 2007).

Örneğin; farklı iki grupta uygulanan bir matematik başarı testinde, eşleme değişkeni olarak sadece toplam test puanı kullanılan FİM analiziyle, farklı işlediği tespit edilen bir soru, okuma becerisi notlarının ikinci eşleme değişkeni olarak kullanıldığı FİM analizinde sorunsuz görünüyorsa, bu sorunun doğru cevaplanmasının, matematik yeterlik düzeyine ek olarak, okuma becerisiyle de ilişkili olduğu yorumu yapılabilir. Bu yorumun bir dayanağı olup olmadığı, çoğunlukla uzman görüşüne başvurularak, ayrıca incelenir. Dolayısıyla, ikinci eşleme değişkeni kullanmak, farklı işleyen soruların farklı işleme sebeplerine ışık tutabileceği için, gruplar arasındaki kültürel farkların belirlenmesine de katkı sağlayabilir. Ancak, bu önemine karşın, FİM analizlerinde kullanılabilecek uygun bir ikinci eşleme değişkeni tespit etmenin, genel geçer bir yöntemi, henüz, belirlenmiş değildir (Gierl 2005).

Bu araştırma, PISA gibi birçok ülkede uygulanan sınavlara yönelik FİM analizlerinde, ikinci eşleme değişkeni olarak kullanılabilecek değişkenleri tespit etmek üzere bir yöntem önermektedir. Yöntem, Grisay ve Monseur'un (2007) çalışmasında kullandıkları faktör çözümlemesine dayanmaktadır. Önerilen yöntemin işe yararlığı, PISA matematik sınav verileri kullanılarak incelenmiştir.

İkinci eşleme değişkenini tespit etmek amacıyla kullanılan faktör çözümlemesinde, gerekli veri, değişkenler ülkelerden, gözlemler ise sorulardan oluşacak şekilde hazırlanmaktadır. Değişkenlerin değerleri, soruların güçlük indisleridir. Bu indisler, her ülke için ayrı ayrı, ilgili ülke öğrencilerinin

sorulara verdikleri cevaplar kullanılarak, Madde Tepki Kuramı'na (MTK) dayalı kestirilmiştir. Çalışmada 40 ülke ve 47 soru kullanılmıştır.

Bu şekilde düzenlenen bir faktör çözümlemesinde, ülkelere ait ortak etken varyans (communality) değerleri, soru güçlüklerindeki varyansın ortak bir faktörce açıklanma oranını vermektedir. PISA matematik sınavında, bu ortak faktör, matematik yeterliği olarak tanımlanmıştır. Dolayısıyla, ortak etken varyans değerinin düşük olduğu ülkelerde, soru güçlüklerinin, matematik yeterliğine ek olarak, başka değişkenlerden de etkilendiği çıkarımı yapılabilmektedir. Bu başka değişkenlerin neler olabileceği korelasyonla incelenmiştir.

Korelasyon çalışması için, PISA öğrenci anketi kullanılarak, her ülke için, öğrencilerin, okul derslerine, okul dışı derslerine ve kendi kendilerine çalışmaya ayırdıkları ortalama süreler hesaplanmıştır. Ülkelere ait bu 3 değişkenle, ülkelerin ortak varyans değerleri arasındaki korelasyonlar incelendiğinde, okul dışı derslere ayrılan ortalama süre ile ortak varyans değerleri arasında negatif yönde bir ilişki bulunmuştur ($r = -.632$, $p < .01$). Bu sonuç, öğrencilerin okul dışı derslere ayırdıkları sürenin, matematik yeterliği dışında, öğrenci performansını etkileyebilecek bir faktör olabileceği şeklinde yorumlanabilir. Bu ihtimal FİM analiziyle incelenmiştir.

Bu amaçla, öğrencilerin ortalama matematik yeterlik düzeyinin birbirine çok yakın olduğu, ancak okul dışı derslere ayrılan süre bakımından oldukça farklı olan Finlandiya ve Kore ülkeleri öğrenci cevapları kullanılmıştır. Sadece, matematik yeterliğinin eşleme değişkeni olarak kullanıldığı FİM analizlerinde, 47 sorudan 10'unun, kullanıldığı her üç kitapçıkta da, Finlandiya ve Kore grupları arasında farklı işlediği tespit edilmiştir. Ancak, öğrencilerin okul dışı derslere ayırdıkları süre, ikinci eşleme değişkeni olarak kullanıldığında, iki sorudaki farklı işleyiş, bütün kitapçıklarda, anlamlı ölçüde azalmış veya tamamen ortadan kalkmıştır. Soruların geçtiği tüm kitapçıklarda benzer sonucun elde edilmesi önemli bir genellenebilirlik kanıtı olarak değerlendirilebilir. Diğer yandan, okul dışı derslere ayrılan süre ile bu sorulardaki performans farklılığı arasındaki ilişki, soruların içerikleri, ülkelerin müfredatları, ilgili ülkelerde okul dışı derslerde yapılan çalışmalar vb. boyutlar dikkate alınarak derinlemesine incelenebilir. Ancak, böyle bir inceleme bu araştırmanın amacı dışında kalmaktadır.

Araştırmada elde edilen sonuçlara dayanarak, bir sınavın uygulandığı ülkelerdeki, madde güçlüklerine ait ortak etken varyans değerleriyle ilişkili olan değişkenlerin, FİM analizleri için uygun bir ikinci bir eşleme değişkeni adayı olabileceği söylenebilir. Bu yöntemle, genelde, ülkedeki madde güçlükleriyle ilişkili olduğu görülen değişkenlerin, özelde, hangi sorulardaki performansı etkilediği incelenebilir.

Araştırmada, madde güçlüklerinin kestirilmesi, faktör çözümlemesi ve FİM analizleri, sırasıyla, Bir Parametreli Lojistik Model, Temel Bileşenler Faktör Çözümlemesi ve Lojistik Regresyon kullanılarak gerçekleştirilmiştir.