ANALYSIS OF DIFFERENTIAL ITEM FUNCTIONING IN COMPARATIVE EDUCATION STUDIES: THE CASE OF THE CIVIC EDUCATION STUDIES

Yuk Fai Cheong
Emory University, USA

Abstract

This study proposes and illustrates a two-stage strategy to analyze differential item functioning in comparative educational studies. At the first stage, within each country, item difficulties are estimated. At the second stage, results are combined across countries using a meta-analytic framework to investigate (a) whether the items function consistently at the country level, and (b) what country characteristics predict item difficulties. To illustrate, we analyzed data on student responses to 13 civic skills items from 28 countries, collected under the Civic Education Study of the International Association for the Evaluation of Educational Achievement (IEA), and reported initial findings. Implications of this strategy for comparative educational studies are discussed.

INTRODUCTION

Valid cross-national comparisons of school achievement require that the assessment items function consistently across different countries or in the adapted- and source-language versions of the test (Hambleton, 2002). A statistical approach to examine the extent of similarity in the behaviors of the items across nations is through the use of a differential item functioning (DIF) study (Clauser & Mazor, 1998). An item is considered to display DIF and pose a potential threat to validity if it is significantly easier or harder for students from different countries, after they have been matched on the proficiency of interest. Psychometricians, for instance, implemented DIF analyses for the Third International Mathematics and Science Study (TIMMS) (Garden & Orpwood, 1996) and the Organization for Economic Co-operation and Development (OECD) Programme for International Student Assessment (PISA) (Adams & Wu, 2002). They used item-by-country interaction statistics, based on the estimates of item difficulty and overall performance, to flag test items for possible DIF associated with country memberships. The items flagged were then reviewed and could be removed from the test, modified appropriately, or administered again.

An important task of reviewing the flagged items for any type of DIF is to identify their possible causes or correlates. The task has great implications for test scoring for different subpopulations of examinees and test construction (Allalouf, Hambleton, & Sireci, 1999). It can provide valuable insights in these attributes of a test and its components: word difficulty, item format, content, and cultural relevance and interpretation (e.g., Allalouf, 2003; Azocar, Arean, Miranda, & Munoz, 2001).

DIF associated with a certain type of membership may mask group differences in educational backgrounds (e.g., Clauser, Nungester, & Swaminathan, 1996), instructional experiences, and opportunities to learn about specific topics in school (e.g., Miller & Linn, 1988; Muthen et al., 1995). For example, Linn and Harnisch (1981) speculated White-African American DIF in achievement items may in fact be caused by differences in the amount of instruction the two groups received. With regard to comparative educational studies, Torney-Purta, Lehmann, Oswald, and Schulze (2001) postulated that cross-national differences in political and historical context and education curricula may bring about the deviation of a country's response pattern from the international findings. A psychometric study of possible causes or correlates of DIF could allow researchers to investigate how the various examinee, school, and country characteristics may relate to student learning and assessment. In this paper, we propose and illustrate an analytical approach that researchers could use to study correlates of country DIF.

The approach combines and analyzes cross-national item difficulties data in two stages. At the first stage, within each country, item difficulties are estimated. At the second stage, results are combined across countries using a meta-analytic framework (Glass, 1976) to investigate (a) whether the items function consistently at the country level, and (b) what country characteristics predict item difficulties. The synthesis to assess the consistency or heterogeneity in item difficulties is analogous to Wright and Stone's (1979, p. 95) separate calibration approach to evaluate the invariance of item difficulty when there are only two countries. Using Wright and Stone's approach, one would perform two independent calibrations, one for each country, and perform t tests to examine for individual items the extent to which the two sets of results estimate the same parameters. An advantage of our approach is that it allows the study of DIF for multiple groups simultaneously. Raudenbush, Cheong, and Fotiu (1995) and Raudenbush, Fotiu, and Cheong (1999) used a similar meta-analytic framework to synthesize cross-national and cross-state achievement and classroom-effects data.

To illustrate, we analyze data on student responses to 13 civic skills items from 28 countries, collected under the Civic Education Study of the International Association for the Evaluation of Educational Achievement (IEA) (Torney-Purta et al., 2001). As suggested by Torney-Purta et al. (2001), in this exploratory analysis, we evaluate whether the social and political context of a country may be related to cross-national differences in item difficulties. Specifically, we examine whether number of Internet hosts and voter-turn-out rate of a country may be related to the deviations in the item response patterns of specific countries from the international results. The former may indicate the level of accessibility to information and the availability of opportunity for civic engagement in a country (Macintosh, Robson, Smith, & Whyte, 2003; Semetko & Krasnoboka, 2003). If associated with item difficulties, we hypothesize that there would be a negative relationship. The latter could indicate the level of political participation, engagement, political self-efficacy as well as trust among voters in a country (Milbrath & Goel, 1982), which could affect the amount of exposure to politically engaged adults the young people have. Again, we hypothesize that it has a negative association with item difficulty.

We also assess whether the difficulty of an item on democracy and its defining characteristics may be related to teacher's perception of student's opportunities to learn about democracy. The focus of the measure is on content coverage (McDonnell, 1995). More exposure is expected to be positively associated with item easiness.

METHODS AND RESULTS

The Data

The illustrative analysis uses the student and teacher survey data collected in 28 countries by the IEA Civic Education Study (Torney-Purta et al., 2001), as well as the cross-national data on human development compiled by the Human Development Report 2000 (United Nations Development Programme, 2000). The IEA Civic Education Study surveyed students and school personnel in 28 countries and its goal was to study and investigate how young people were prepared for citizenship in democracies. Its two foci were the in- and out-of-school experiences relevant to the initiation of young people into political communities.

Instruments and Measures

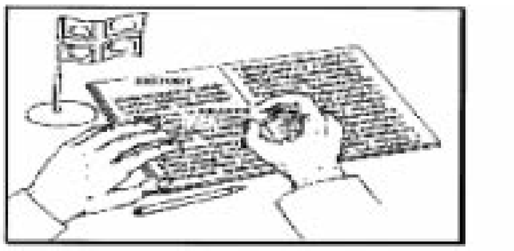
The IEA Civic Education Study employed item response theory methods to develop 13 multiple-choice items assessing students' skills in using civic-related knowledge (Baldi, Perie, Skidmore, Greenberg, & Hahn, 2001; Torney-Purta et al., 2001).The domain content categories covered by the items included democracy and its defining characteristics, institutions and practices in democracy, rights and duties of citizenship, national identity, international relations, and social cohesion and diversity. The IEA Study provided short titles for all items and released the content of three of them. Figure 1 displays the content of the released items.

Figure 1. IEA Civic Education Study Released Civic Skills Items

> We citizens have had enough!
> A vote for the Silver Party means a vote for higher taxes.
> It means an end to economic growth and a waste of our nation's resources.
> Vote instead for economic growth and free enterprise.
> Vote for more money left in eveyrone's wallet!
> Let's not waste another 4 years! VOTE FOR THE GOLD PARTY.

1. This is a political advertisement that has probably been issued by...
a. The Silver Party.
b. A party or group running against the Silver Party.
c. A group which tries to be sure elections are fair.
d. The Silver Party and the Gold Party together.

2. What is the message or main point of this cartoon?



History textbooks...
a. are sometimes changed to avoid mentioning problematic events from the past.
b. for children must be shorter than books written for adults.
c. are full of information that is not interesting.
d. should be written using a computer and not a pencil.

3. Three of these statements are opinions and one is a fact. Which of the following is a FACT?
a. People with very low incomes should not pay any taxes.
b. In many countries rich people pay higher taxes than poor people.
c. It is fair that some citizens pay higher taxes than others.
d. Donations to charity are the best way to reduce differences between rich and poor.

Data Analytic Strategy

In the first stage or within-country analysis, we estimate the difficulty of the items for individual countries. We employ a Rasch model (Wright & Stone, 1979) for the evaluation of model fits and estimation. The output from the within-country analysis consists of estimates of item difficulties in logits and its estimated sampling variance. The second stage or between-country analysis combines the output produced by each country to obtain inferences on all model parameters. It involves two analyses: assessing and accounting for heterogeneity in country item difficulties. We use an empirical Bayes approach to synthesize the country-by-country results for the models. The following equations express the models estimated to assess heterogeneity in country item difficulties:

Table 1 gives the descriptive statistics for the country-level predictor variables.

Table 1: Descriptive Statistics for Country-Level Variables

Country	*Internet Hosts (per 1000 people)*[a]	*Voter Turn-out at Latest Elections*[b]	*% of students whose Teachers Perceive "considerable" or "Very Much" Opportunity to Learn About Democracyc*
Australia	41.0	95	28
Belgium	20.6	91	18
Bulgaria	1.2	68	26
Chile	2.0	86	32
Colombia	0.4	45	n.a.
Cyprus	7.9	93	26
Czech Republic	8.4	74	51
Denmark	56.3	86	77
England	24.6	72	25
Estonia	16.6	57	21
Finland	89.2	65	36
Germany	17.7	82	18
Greece	4.7	76	27
Hong Kong	12.4	43	13
Hungary	9.4	56	30
Italy	6.7	83	42
Latvia	5.8	72	37
Lithuania	2.7	53	15
Norway	71.8	78	47
Poland	3.4	48	61
Portugal	5.6	62	34
Romania	1.1	76	34
Russia Federation	1.2	62	74
Slovak Republic	4.1	84	87
Slovenia	11.5	74	57
Sweden	42.9	81	87
Switzerland	34.5	43	33
United States	112.8	36	n.a.
Mean	21.99	69.32	39.85
Standard deviation	28.81	16.11	21.70

Notes: a, b Source: National Development Report 2000 (United Nations Development Programme, 2000).

c Data from Columbia and the United States are omitted due to country-specific problems in ascertaining the linkage between teachers and classes of students.

$$\begin{aligned} b_{pk} &= \beta_{pk} + e_{pk},\ e_{pk} \square N\ (0,\ v_{pk}) \\ \beta_{pk} &= \gamma_p + \mu_{pk},\ \mu_{pk} \square N\ (0,\ \tau_p) \end{aligned} \qquad (1)$$

In the model, b_{pk} is the difficulty estimate for item p in country k, $p = 1, \ldots, 13$, $k = 1, \ldots, 28$. The estimate b_{pk} is assumed to vary around its corresponding parameter β_{pk} with a unique error associated with the sample for country k, e_{pk}, which has a known sampling variance v_{pk}. The parameter β_{pk} is in turn assumed to vary around an overall mean γ_p plus a random error associated with country k, μ_{pk}. The random error has a variance of τ_p. To account for the variability, we regress β_{pk} on the three predictors, Number of Internet Hosts, Voter Turn-Out, and Opportunity to Learn. To reduce the skewness of the distribution of Number of Internet Hosts, we took a logarithmic transformation of the variable.

Results

Results for the individual countries from the first stage analysis showed that all the items had acceptable fits. For each country, 13 $\hat{b}_{pk}$'s were output for the second stage analysis. Table 2 reports the estimates of γ_p and their associated τ_p, grouped by their domain content categories.

Results indicated that the hardest item across nations was Item 7 in the Rights and Duties of Citizenship domain, $\hat{\gamma}_7 = 1.519$ and the easiest item was Item 13 in the Social Cohesion and Diversity domain, $\hat{\gamma}_{13} = -.704$. Item 2 displayed the greatest variability, $\hat{\tau}_{13} = .310$. Findings from the tests of heterogeneity of country mean item difficulties suggested statistical significance. They recommended use of caution in any cross-national comparisons based on these civic skills items.

We used Number of Internet Hosts and Voter Turn-out as predictors to account for the heterogeneity in country item difficulties. Initial results showed that the two predictors were related to the difficulties of Item 12 in the Social Cohesion and Diversity domain and Item 9 in the National Identity domain. Controlling for Voter Turn-out, Number of Internet Hosts predicted the difficulty of Item 12 (Recognize groups subject to discrimination). The estimate of the effect was -.241, *s.e.* = .056, $t = -4.343$, $p = .000$. As hypothesized, an increase in the number of Internet hosts was associated with less item difficulty. In terms of effect size, one standard deviation increase in Number of Internet Hosts was associated with a change of -.350 logits in difficulty. Voter Turn-out was not associated with the difficulty of Item 12. The estimate of the effect was .009, *s.e.* = .005, $t = 1.861$, $p = .074$.

Controlling for Internet Host, Voter Turn-out was negatively associated with the difficulty of Item 9 (Recognize sense of collective identity). The estimate of the effect was -.011, *s.e.* = .004, $t = -2.812$, $p = .010$. As hypothesized, a greater percentage of voter turn-out was associated with less item difficulty. In terms of effect size, one standard deviation increase in Voter Turn-out was associated with a change of -.177 logits in item difficulty. After controlling for Voter Turn-out, Number of Internet Hosts did not predict the difficulty of Item 9. The estimate of the effect was .059, *s.e.* = .007, $t = 1.295$, $p = .207$.

Table 2: Results for the Between-Country Synthesis

Domain Content Category	*Item*	*Approximate Posterior mean of γ_ρ*	*Estimate of between-country variance τ_ρ*
I A: Democracy and its defining characteristics			
Evaluate strength and weakness of democratic system	1	-.020	.126
I B: Institutions and Practices in Democracy			
Identifying qualifications of candidates for positions and making up one's mind during election	2 3 4	-.323 -.700 .178	.310 .120 .213
Identifying a healthy critical attitude toward officials and their accountability	5	-1.101	.138
Understand basic economic issues and their political implications	6	.709	.105
I C: Citizenship: Rights and Duties			
Identify network of associations and differences of political action	7	1.519	.264
Demonstrate awareness of tradeoffs	8	-.358	.275
IIA: National Identity			
Recognize sense of collective identity	9	-.368	.147
Recognize that every nation has events in its history of which it is not proud	10	.168	.251
II B: International Relations			
Recognize international economic issues and organizations (other than inter-governmental) active in dealing with matters with economic implications	11	.407	.154
III A: Social Cohesion and Diversity			
Recognize groups subject to discrimination	12 13	.593 -.704	.295 .127

Finally, we evaluated if Opportunity to Learn about Democracy might predict the difficulty of Item 1 in the Democracy and its Defining Characteristics domain (Evaluate strength and weakness of democratic system) for 26 out of the 28 countries. Data from Columbia and the United States were omitted due to problems in matching the teachers to classes of students (Torney-Purta et al., 2001). The findings indicated that the variable did not predict the difficulty of Item 1. The estimate of the effect was -.002, *s.e.* = .003, t = -.018, p = .540.

SUMMARY AND CONCLUSION

In this analysis, we explored a two-stage approach to synthesize cross-national item difficulties data to assess and account for heterogeneity in country item difficulties. In the first stage, within each country, item difficulties are estimated. At the second stage, results are combined across countries using a meta-analytic framework to investigate (a) whether the items function consistently at the country level, and (b) what country characteristics predict item difficulties. We illustrated the strategy using a small data set from the IEA Civic Education Study. The results indicated significant heterogeneity in country item difficulties for all items. Initial findings showed that Number of Internet Hosts and Voter Turn-out Rate were negatively associated with the difficulties of two of the 13 items on interpretative civic skills. Opportunity to Learn about Democracy, however, did not predict the difficulty of an item on democracy.

This exploratory study shows that the synthesis of cross-country item difficulties could offer us opportunities to learn how country characteristics relate to student learning and assessment. For instance, the initial findings obtained may suggest that valid cross-national comparisons of the proficiency tapped by Items 9 and 12 would require consideration of and adjustments for the two predictors. They also indicated that the two items might have measured the interrelated dimensions represented by the predictors, such as the opportunity for political engagement via Internet or the exposure to politically engaged and self-efficacious adults. The empirical associations of these dimensions with DIF may also provide information on the conceptualization of the process of political socialization. Besides achievement, classroom- and school-effects data for individual countries, international educational surveys could provide yet another cross-national lenses into the development of students through their data on country item difficulties.

References

Adams, R. J., & Wu, M. (Eds.). (2002). *PISA 2000 technical report.* Paris, France: OECD Publications.

Allalouf, A. (2003). Revising translated differential item functioning items as a tool for improving cross-lingual assessment. *Applied Measurement in Education, 16*, 55-73.

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

Azocar, F., Arean, P., Miranda, J., & Munoz, R. F. (2001). Differential item functioning in a Spanish translation of the Beck Depression Inventory. *Journal of Clinical Psychology, 57,* 355-365.

Baldi, S., Perie, M., Skidmore, D., Greenberg, E., & Hahn, C. (2001). What democracy means to ninth-graders: U.S. results from the International IEA *Civic Education Study*. NCES 2001-096. Project Officer: Dawn Nelson. Washington, DC.: Department of Education, National Center for Education Statistics.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. An NCME instructional module. *Educational Measurement: Issues & Practice*, *17*, 31-44.

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*, 453-464.

Garden, R. A., & Orpwood, G. (1996). Development of the TIMSS achievement tests. In M. O. Martin & D. L. Kelly (Eds.), *Third International Mathematics and Science Study (TIMSS) technical report, volume I: Design and development.* (pp. 2-1 to 2-20). Chestnut Hill, MA: Boston College.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*, 3-8.

Hambleton, R. K. (2002). Adapting achievement tests into multiple languages for international assessments. In A. C. Porter & A. Gamoran (Eds.), *Methodological advances in cross-national surveys of education achievement* (pp. 58-79). Washington, DC: National Academy Press.

United Nations Development Programme (2000). *Human development report 2000.* Oxford: Oxford University Press.

Linn, R. L., & Harnisch, D. L. (1981). Interactions between item content and group membership on achievement test items. *Journal of Educational Measurement*, *18,* 109-118.

Macintosh, A., Robson, E., Smith, E., & Whyte, A. (2003). Electronic democracy and young people. *Social Science Computer Review, 21*, 43-54.

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

Milbrath, L. W., & Goel, M. L. (1982). *Political participation; how and why do people get involved in politics?* (2nd ed.). Chicago: Rand McNally.

Miller, M. D., & Linn, R. L. (1988). Invariance of item characteristic functions with variations in instructional coverage. *Journal of Educational Measurement*, *25*, 205-219.

Muthen, B. O., Huang, L.-C., Jo, B., Khoo, S.-T., Goff, G. N., Novak, J. R., & Shih, J. C. (1995). Opportunity-to-lean effects on achievement: Analytical aspects. *Educational Evaluation and Policy Analysis*, *17*, 371-403.

Raudenbush, S. W., Cheong, Y. F., & Fotiu, R. P. (1995). Synthesizing cross-national classroom effect data: Alternative models and methods. In M. Binkley & K. Rust & M. Winglee (Eds.), *Methodological issues in comparative international studies: The case of reading literacy.* (pp. 243-286). Washington: DC.: National Center for Educational Statistics.

Raudenbush, S. W., Fotiu, R. P., & Cheong, Y. F. (1999). Synthesizing results from the Trial State Assessment. *Journal of Educational and Behavioral Statistics*, *24*, 413-438.

Semetko, H. A., & Krasnoboka, N. (2003). The political role of the Internet in societies in transition - Russia and Ukraine compared. *Party Politics, 9*, 77-104.

Torney-Purta, J., Lehmann, R., Oswald, H., & , & Schulz, W. (2001). *Citizenship and education in twenty eight countries: Civic knowledge and engagement at age fourteen.* Amsterdam: The International Association for the Evaluation of Educational Achievement.

Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch Measurement.* Chicago: MESA Press.