Cognitive interviewing evidence on DIF in Polytomous Items of the Student Questionnaire of the PISA

Jose-Luis Padilla (jpadilla@ugr.es)

University of Granada (Spain)

Isabel Benítez Baena (ibenitez@ugr.es)

University of Granada (Spain)

M. Dolores Hidalgo (mdhidalg@um.es)

University of Murcia (Spain)

Stephen G. Sireci (sireci@acad.umass.edu)

University of Massachusetts Amherst (USA)

Paper presented at the 42th Annual Conference of the Northeastern Educational Research Association. Rocky Hill, CT (USA). October,19-21, 2011.

Abstract

Research on Differential Item Functioning (DIF) in international and cross-cultural assessments has focused much more on developing statistics than on explaining DIF results. Qualitative evidence from cognitive interviewing may be helpful in explaining DIF results. This study is part of a major research project aimed at investigating the causes of DIF by means of cognitive interviewing. Starting from DIF results obtained by analyzing US and Spanish versions of the Student Questionnaire of the Program for International Student Assessment (PISA, OECD, 2006), 20 cognitive interviews in the US and 24 in Spain were conducted. Interview protocols were developed taking expert appraisal evidence into account. Interviewees respond to general and follow-up probes after answering each Student Questionnaire scales. Preliminary cognitive interviewing evidence on some of the 8 items flagged with large DIF is presented. Lastly, arguments for extending the use of cognitive interviewing to explain DIF results will be pointed out.

Keywords**:** DIF causes, cognitive interviewing, validity evidence, PISA study.

Cognitive interviewing evidence on DIF in Polytomous Items of the Student Questionnaire of the PISA

Cross-cultural testing has become one of the most important topics in educational and psychological research. Projects like the Programme for the International Assessment of Adult Competencies, PIAAC (Organisation for Economic Co-operation and Development, 2004) and the Program for International Student Assessment, PISA (Organisation for Economic Co-operation and Development, 2006), are just two examples of international programs that regularly evaluate and compare people around the world.

Despite the fact that important decisions are made based on comparing countries, such comparisons are sometimes made without taking the cultural and linguistic differences into account. As the International Test Commission (ITC) and others have pointed out, when translated versions of assessment instruments are used to compare groups and individuals, the consistency of measurement across languages must be established (Hambleton, Merenda, & Spielberger, 2005; International Test Commission, 2010).

One useful strategy for evaluating the equivalence of items that have been translated for use in cross-cultural assessment is analyzing differential item functioning (DIF). DIF occurs when examinees with the same proficiency level on the characteristic or attribute measured, but who belong to different groups (demographic, linguistic, or cultural), have a different probability of giving a specific item response (Millsap & Everson, 1993). DIF analyses identify items that function differentially so that these items can be inspected to determine whether the difference may be due to some form of construct-irrelevant variance. In the case of cross-lingual assessment, one potential

cause of DIF is a translation problem. That is, the meaning of an item could be altered via the translation process (Sireci, Patsula, & Hambleton, 2005). Thus, DIF analysis is an important tool for evaluating the validity of instruments used in international comparisons.

Although detecting DIF can provide relevant validity evidence, research efforts have been focused much more on developing statistics to flag item with DIF than on identifying DIF causes. In relation with the DIF causes research, Padilla, Pérez and Pérez (1998) pointed out three points which could summarize main results obtained until that moment: a) the effect of structural characteristic of the items could be explained by differences in cultural and formative experiences of the groups involved (e.g., Schmit & Dorans, 1990); b) traditional socio-demographics variables used to define the groups could be hiding relevant factors to explain DIF (e.g., Muthén, 1988); and c) the benefits of defining groups by hypothetical relevant factor for explaining DIF instead of traditional demographic variables (e.g., Miller & Linn, 1988). Even though DIF research has grown at the same time as interest in adapting tests and questionnaires, evidence on DIF causes are still scants (Allalouf, Hamblenton, & Sireci, 1999; Ferne & Rupp, 2007; Sireci, 1997).

This study is part of a major research project aimed at investigating the causes of DIF by means of cognitive interviewing. The rationale behind of the research project is that it is necessary to understand the difference in the response process to the DIF items to answer the question of why DIF appears (Ercikan, Arim, Law, Domene, Gagnon and Lacroix, 2010).

The *Standards* (AERA, APA, NCME, 1999) point out "many lines of evidence can contribute to an understanding of the construct meaning of test scores" (p. 5) and suggests that those lines of evidence can consist of the familiar categories of evidence

based on test content, response processes, internal structure, relations to other variables, and consequences. "Evidence based on the response processes" resort to the empirical and theoretical analysis of the response processes of the respondents in order to obtain evidence about the fit between the construct which test or questionnaire intend to measure and the response actually put in to practice by examinees. According to the *Standards*:

> *"Evidence based on response processes generally comes from analyses of individual responses. Questioning test takers about their performance strategies or responses to particular items can yield evidence that enriches the definition of the construct" (p.12).*

The proposal on which the present study is based is that cognitive interviewing can provide evidence "based on the response processes" of the respondents useful to explain DIF results in cross-cultural and national testing.

The use of cognitive interviewing as a method for evaluating the quality of questions included in surveys has become more and more popular (Castillo, Padilla, Gómez-Benito, and Andrés, 2010). On general terms, cognitive interviewing is a semi-structured interview that seeks to gather information about the "question/item-and-answer process" that respondent has performed when answering the questionnaire by applying general and follow-up probes (Beatty and Willis, 2007). Cognitive interviewing provides evidence about problems with the comprehension of key terms, failures in the data collection, errors in the wording of the question and mismatch in the choice of response option designed for the questionnaire. Although the application of cognitive interviewing has been mainly related to questionnaires included in surveys, it can also be implemented for the assessment of psychological scales.

With the aim of investigating into the usefulness of cognitive interviewing to explain DIF results, some scales included in PISA (Organization for Economic Co-

operation and Development, 2006) were used. The Program for International Student Assessment (OECD, 2006) is a dynamic model widely used that evaluate the so-called cross-curricular competencies as students' approaches to learning, what includes the way students address and handle learning tasks in school and the extent to which they are able to achieve their learning goals by applying strategies, motivating themselves, and by controlling and regulating their own learning processes. In this type of studies, the existence of DIF can undermine conclusions about differences between groups. But more than detecting DIF, identifying the causes may provide relevant validity evidence.

After analyzing US and Spanish versions of the Student Questionnaire of the Program for International Student Assessment (PISA, OECD, 2006), the objective of this study is to interpret DIF results by means of cognitive interviewing. First, the main points of the DIF study will be summarized. Secondly, the cognitive interviewing study will be presented. Finally, some issues related how to combine qualitative and quantitative evidence to improve educational measurement will be discussed.

Study 1: Detecting Polytomous DIF in PISA 2006 attitudinal items.

The aim of this study was to detect polytomous DIF by Mantel-Haenzsel and Ordinal Logistic Regression (OLR) procedures. DIF across English and Spanish versions of the seven scales included in the Student Questionnaire of the Program for International Student Assessment (PISA, OECD, 2006), was analyzed.

Method

Participants

In this study, responses of 17,405 participants from Spain (8,704 women and 8,701 men) and 4,902 participants from the United States (2,422 women and 2,480 men) were analyzed. The participants from Spain were all 16 years old and those from the US were between 15 and 16 years old (mean 15.5 and SD 0.5).

Instruments

Seven scales were selected for analysis from the PISA student questionnaire. All were 4-point Likert item scales intended to measure science related attitudes. PISA divided the scales into four main topics: interest, support, motivation to learn, and self-cognitions. Table 1 presents the scale topic, the intended construct for each scale (and the short which will be used across the paper for identifying the items in each scale), the number of items in each scale and the name assigned to the scale in the PISA dataset.

Table 1

Characteristics of PISA scales used in the DIF analyses.

Topic	Construct	# of Items	Dataset Name
Interest of science learning	Enjoyment of science (Enj)	5	JOYSCIE
Support with science	General value of science (Gen)	5	GENSCIE
	Personal value of science (Per)	5	PERSCIE
Motivation to learn science	Instrumental motivation to learn science (Ins)	4	INSTSCIE
	Future-oriented to science motivation (Fut)	5	SCIEFUT
	Science self-efficacy (Eff)	8	SCIEEFF
Self-related cognitions in science	Science self-concept (Con)	6	SCSCIE

The selection of the scales was based on two criteria. First, the selected scales should have been developed as a unidimensional scale; and secondly, PISA researchers use the total scale score in a descriptive (e.g., to describe student attitudes to science) or statistical (e.g., in the computation of students' plausible value scores) manner.

Procedure

Data were obtained from the OECD website (http://www.oecd.org). Scales and subjects were selected and data cleaning was conducted by eliminating the subjects with incomplete responses to one or more survey questions. This process resulted in a relatively minor loss of data (12.6% and 11.2% for the USA and Spanish samples, respectively). Subsamples (described next) were obtained from this cleaned dataset. Finally, DIF analyses were computed and the reliability of the results was checked by comparing the results across two independent replications. DIF analyses were done using country as the group variable. With the aim of having comparable group sizes and replication to confirm any statistical conclusions, two random subsamples of 4,900 Spanish participants were selected. Both Spanish subsamples were compared with the US group.

Analyses

Due to the polytomous nature of the survey items DIF was analyzed using Penfield's *differential step functioning* (DSF) framework (Penfield, 2007, 2010; Penfield, Gattamorta, & Childs, 2009). To conduct the DSF analyses, we used the DIFAS 4.0 software (Penfield, 2007), which evaluates DIF/DSF using the odds ratio approach to test the null hypothesis of no DSF. We used DIFAS to first analyze overall DIF (i.e., DIF at the item level), and then subsequently to evaluate DSF in items that were flagged as showing overall DIF. The Standardized Liu-Agresti Cummulative Common Log-Odds Ratio (LOR Z) was used to flag items for DIF, in which a value greater than 2.0 or less than –2.0 is considered evidence of the presence of DIF (Penfield & Algina, 2003). DSF analysis was applied for items flagged with DIF in both subsamples applying cumulative categories with three steps (since there were four response categories in each item).The effect size for evaluating DIF items component was $\hat{\lambda}_j$ (the step-level log-odds ratio estimator), with | ˆλ j | < .43 signifying a small or

negligible effect, $.43 \leq |\hat{\lambda}_j| < .64$ signifying a medium effect, and $|\hat{\lambda}_j| \geq .64$ signifying a large effect (Penfield, 2009). Items detected in both subsamples with medium or large effect were considered to exhibit DIF.

On the other hand, OLR analyses were performed with the Statistical Package for Social Sciences (SPSS v.16) by following the instructions elaborated by Zumbo (1999). First, items were analyzed in both replications obtaining a chi-square significance value which was used for determining items with DIF. Later, log-OR-GR values and delta index values were checked in items detected with DIF in both samples, and DIF effect size was classified. The effect size classification was done by following Penfield (2009) criterion explained above for log-OR-GR and using the Educational Testing Service (ETS) criterion for interpreting Delta index. In this case DIF was considered as medium when values between 1 and 1.5 and large when values higher than 1.5.

Results

Due to the aim of this paper is to illustrate how cognitive interviewing can help in interpreting DIF results, in this section we only present the main results of the DIF analyses, specifically, the convergence across the MH and OLR results. Additional details can be provided on request.

DIF results: Convergence across methods

Before conducting DIF analyses, the dimensionality of responses to items in both groups was analysed using exploratory factor analysis (principal axis method). Separate analyses were done for each scale. In all the groups a dominant single factor was obtained, with the first factor accounting for at least 46% of the variance in the data for all the scales (VAF ranged from 46% for the Science Self-Efficacy scale in Spanish

sample 2, to 80% for the Motivation to Learn Science scale in Spanish sample 1). These values confirm the unidimensionality of the scale according to typical criteria established in the literature (Carmines and Zeller, 1979; Reckase, 1979).

DIF analyses were computed for all 38 items across the seven scales. The analyses were done separately for each scale, thus, the matching criterion was the total score across the subset of items comprising each scale. 20 items were marked as exhibiting in both samples medium or large DIF by MH. Of them, 17 had the same effect classifications in both replication (considering only medium and large effects) and three of them (Eff2, Gen2 and Per1) registered different effect size in both cases. 24 items were detected by ORL procedure as exhibiting medium or large DIF considering both criterions and both replications. In all the items, the effect size estimated was consistent. Inconsistencies were found but not in these 24, and it occurred only in two items (Gen4 and Con5) in which medium sizes were obtained for all the cases except for ETS criterion in replication 2.

For the purpose of the this research, items considered finally as items with DIF were those flagged with both procedures in both samples with medium or large DIF. Using this criterion it is possible to guarantee with more security items flagged really have DIF. For establishing the items with DIF, first effect size comparisons across replications were done separately for each method. In OLR also convergence across criterions in each replication were compared. Later, only items obtained from both procedures were selected and effects sizes were studied. For facilitating the items classification, a convergence table was done by considering the final results from comparing methods. Table 5 shows the results classifying items based on the agreement across methods. Items are identified by the scale short specified in table 1 and the item number in the scale.

Table 2

Convergence across methods

		OLR	
		Large	Medium
MH	Large	Enj5, Eff8,Gen1, Gen3, Gen5, Per5, Ins1, Con4	Enj3, Eff1, Eff5
	Medium	Per4, Ins4, Con3	

Table 2 shows the agreement across methods for classifying effect size. It all the items included in the table, results were equal for both replications and for all the criterions included in each method, it is, it was found within-method consistency. Only in the case of small size, inconsistency was found across both criterions in one of the replications. It occurred in items Gen4 and Con5 were effect size was classified as medium in the first replication (both criterion) and log-OR-GR criterion in the second replication, but it was classified as small when using the ETS criterion in the second replication. As table 5 shows, in eight items a complete agreement was found across methods and samples- These items were those classified as having large DIF in both analyses. On the other hand, six items were flagged with both methods but with a different effect size.

Study 2: Obtaining validity evidence by cognitive interviewing to interpret DIF in PISA 2006 attitudinal items.

As it was stated before, the main objective of this study was to determine whether the evidence provided by cognitive interviews facilitates the interpretation of the DIF results obtained from statistics. The analysis of these response processes for US and Spanish respondent could allow us to investigate potentially different

interpretations of the intended construct. Then, we intended to relate the evidence provided by cognitive interviewing with the DIF results. Two main research questions guided analyses: a) do Spanish and US participants interpret differently item contents (concepts, key terms, etc.)?; and b) if they do, can theses differences be associated with DIF results?

Method

Participants

20 interviews were conducted in the States and 24 in Spain. Table 3 show the main demographics for participants in both groups. All participants were 15 or 16 years old because both ages are the target ages for the PISA study.

Table 3. Demographics of participants in cognitive interviewing

	Gender		**Age**	
	Male	**Female**	**15**	**16**
USA (20)	9	11	11	9
Spain (24)	9	15	9	15

18 of the 20 US participants were enrolled in Grade 10 or 11 when interviews were conducted, while 19 of 24 Spanish participants were in the equivalent Spanish grades. All US participants are native English speaking and all the Spanish participants speak Spanish as theirs first language.

US interviews took place in Chicago and six suburbs ranging from the far south suburban Chicago area to northern suburbs. Five US respondents were recruited using a local youth job center, the rest through word of mouth. There was one private school student; the rest went to local public schools. Schools included a wide range of socio-economic and ethnic diversity. Two interviews were conducted in a public library study room, four at local coffee shops and the rest in private homes.

Spanish interview were conducted in Granada. All respondent were contacted through school principals and word of mouth. There were 12 private school students, and 12 students that attended a local public school. The public school includes a wide range of socio-economic diversity; while the private school includes middle and middle-high socio-economic status.

Materials

Interviewers in both countries used an interview protocol which included general and follow-up probes. Interview protocols were developed taking expert appraisal evidence into account. 11 experts with strong background in education, test developments and cross-cultural testing were asked to rate to what extent US English and Spanish items are comparable. In addition, experts were encouraged to provide comments on linguistic issues: terms, expressions, etc. that could undermine item comparability. General and follow-up probes were included in the protocol for the 8 items flagged with large DIF in Study 1 and items that experts pointed out potential problems of comparability. The protocol were developed in Spanish and translated into English by a committee-approach design. Table 4 shows the section of interview protocol for the Q16 Student Questionnaire scales.

Table 4. Example of general and specific follow-up probes for Q16 scale (“Enjoyment of Science”)

P.16. 1. Let's start talking about how you answered the first questions. The first questions were about "Your opinion about science." They are questions about whether you have fun learning "scientific issues", if you enjoy, etc., learning "things about science", etc., how you have understood these questions, in other words, what you thought while answering these questions?

⇒ SPECIFICS:

P. 16. 2. Tell me what “broad science topics” etc. are for you in school.
P. 16. 3. When you responded to the first statement which said “I generally have fun when I am learning broad science topics”, what situations were you thinking of? (places, times, etc.)
P. 16. 4. Also, statement b) says “I like reading about broad science”. What situations were you thinking of when responding?
P. 16. 5. Tell me examples of “broad science problems” you have thought about when responding to the statement “I am happy doing broad science problems.”
P. 16. 7. In the phrase "I am interested in learning about broad science ", you have answered _____ (See and read the alternative marked by the participant in statement e), explain your answer, why did you answer that.
P. 16. 8. What have you understood for “I am interested”, in the sentence “I am interested in learning about broad science”?
P. 16. 9. Your reply has been ____ (See and read again the alternative marked in statement e) In this sentence, what would your "interest in learning about broad science" be? (what would you do, think, etc..) so that your answer would be ____ (read the alternative furthest from that marked by the participant).

In addition, official versions of the Spanish and US Student Questionnaires were used.

Procedure

The follow-up probes included in the protocol were applied retrospectively, i.e. first participants responded to the whole PISA Student Questionnaire, and then the general and follow-up probes were applied. The retrospective application of the probes is appropriate when the presentation of the items is desired to be as realistic as possible (Willis, 2005). In this case, the retrospective application enabled the administration of the PISA Student Questionnaire carried out in Studies 1 and 2 to be comparable.

Two US interviews were conducted in a public library study room, four at local coffee shops and the rest in private homes. Spanish interviews took place in school offices.

Analyses

The analysis of the cognitive interview data was conducted following the approach, in several stages, developed by Miller (2007). In the first stage, the interviews were analysed transcripts individually in order to reveal the participants' interpretations of the item contents. Analysts used Q-Notes, an on-line data entry and analysis software application developed at *National Center for Health Statistics* (https://wwwn.cdc.gov/qnotes/login.aspx). From this data set main themes were established, within which the participants developed different sub-themes in both groups. During the second stage, the interpretations made by different groups of participants defined by country were compared. This comparison made it possible to test whether the problems with the scale items were specific to a group or common to all participants. Differences between groups were analysed based on the interpretations of the indicators developed in the replies of the participants.

Ethics and data collection

First, participants were informed about the purpose of the study. To motivate participants, they were told how important the interviews will be to improve a cross-national study sponsored by the University of Massachusetts Amherst and the University of Granada in Spain. The interviewers told the participants the information provided by the study will be used by policy makers and researchers to improve education. In addition, each participant was rewarded with a fancy stick memory. The interviews were conducted individually by four trained and experienced interviewers (three females and one male). The interviews were recorded on audio with the consent

of the participants' parents. The participants and theirs parents were guaranteed confidentiality and that the data would be solely used for purposes related to research.

Findings

To illustrate how cognitive interviewing can help in understanding DIF results, in this section we only present findings for item Q16.1 ("*I generally have fun when I am learning broad science topics*"), and item Q16.5 ("*I am interested in learning about broad science*"). Item Q16.1 didn't show DIF while item Q16.5 were flagged with large DIF for MH and ORL procedures.

The evidence obtained from the cognitive interviews is presented in two parts. The first part shows the themes developed by the participants for both items. The second part presents the differences detected between the interpretations made by Spanish and US participants.

The general and specific follow-up probes included in the interview protocol obtained from the participant narratives themes and sub-themes that were identified by analysts. Table 5 shows the frequencies of themes and sub-themes identified from the participant narratives in both country groups:

Table 5. Themes and sub-themes for item Q16.1

Theme	**Subtheme**	**Spain**	**USA**
Broad science	School subjects	20	13
	Not School subjects	12	9
Situations	Classes	18	17
	Study/homework	6	3

	Other materials	3	4

Example 1 shows some of the statements of the Spanish and US participants about the sub-theme "school subject" identified to understand the meaning of "science":

> ["Lo que aprendemos en las asignaturas de ciencias que aplicamos a la vida normal… es que hay muchas cosas… he pensado por ejemplo si me enseñan una cosa en física y química o en matemáticas y en si me va a servir para otras cosas, y así amplio mis conocimientos, un poco de cultura general". S04]
>
> [(English version): "… what we learn in scientific subjects that we use in normal life… there's lots of things… for example, I thought of they teach me a issue in physic and chemistry or math and if these things will be useful for me to do other things, increasing my knowledge, a bit of general culture". S04]
>
> ["Like the Science that you typically learn in school like the Biology, and Chemistry, and the Physics, things like that, Computer Sciences also". US01]
>
> ["I think of all the labs I did, in biology and chemistry and now physics". US15]

Frequencies and participant narratives allow us to infer that the understanding of the key concepts in item Q16.1 was very similar in both groups.

Table 6 show the frequencies of themes and sub-themes identified by analyst to find out the understanding of key concepts in item Q16.5.

Table 6. Themes and sub-themes for item Q16.5

Theme	**Sub-theme**	**Spain**	**USA**
I am interested	Interested in the topics	15	20
	Advisable (good for me)	10	0

Example 2 shows some of the statements of the Spanish and US participants about the sub-theme "interested in the topics" identified to understand the meaning of "I'm interested":

> ["significa que tengo interés por aprender eso… que tengo ganas de aprender sobre la ciencia" S05]
>
> [(English version) "… that mean I'm interested in learning about that… I feel like learning about science" S05]
>
> ["Like, it's something that you'd actually like to do or watch someone do. Like, it's something that, you want to learn more about it and you, you're really into it, it's something that you like". USA04]
>
> ["Interested in like you enjoy it kind of, more than just having to learn about it. Well I pay attention and I listen during class and stuff. Yeah. I do an oceanography program during the summer so it's kind of like I'm interested about that. So I guess that's learning about science". USA07]

None of US participants gave narratives of the sub-theme "Advisable (good for me)", while 10 Spanish participants talked about that when responded to the specific follow-up probes. Example 3 shows some of the statements of the Spanish participants:

> ["Estoy pensando en el futuro, porque me interesa aprender cosas sobre ciencias porque me viene bien si voy a hacer una carrera de ciencias, lo he entendido como que me conviene para el futuro". S06]
>
> [(English version) "I'm thinking of the future, because it is advisable for me to learn things about sciences … I'm going to enroll in science grade, I understood that it's good for me in the future". S06]
>
> ["Que para mí me interesa porque es algo que me va a servir en un futuro, sino me interesase porque yo creo que no va a ser útil para un futuro pues no… una cosa es que te guste y otra cosa es que yo crea que me va a servir". S12]
>
> [(English version) "It's advisable for me because it's something that will be useful in the future… If I'm not interested because I believe that it won't be useful in the future… one thing is that you like it and other that I think it will be useful" S12]

As none of the US participants gave narrative including the subtheme "advisable" ("good for me in the future"), while 10 Spanish participant did, it is possible that US and Spanish participants understand differently one of the key concept in item Q16.5.

Discussion

Despite of the results of Study 2 are preliminary, we think they are very promising. The analyses of the US and Spanish participants show different interpretation patterns for most of the items flagged with large DIF in Study 1. The differences in the interpretation patterns affect the understanding of key terms in DIF items. The differences are related to different schooling experiences, educational contexts, and, in some cases, term and expression with different meaning in both languages.

Research on linking differences in the interpretation patterns to differences in response patterns is being conducted. In addition, comparisons of the narratives of matched Spanish and US participants using scale total scores are being performed. These comparisons are needed to connect cognitive findings and DIF results. Lastly, we are analyzing evidence to know to what extent responses and demographics of the participants in cognitive interviewing are comparable to those of the PISA 2006 Spanish and US respondents.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. Washington, D.C.: American Psychological Association.

Allalouf, A.; Hamblenton, R.K. & Sireci, S. (1999). Identifying the Causes of DIF in Translated Verbal Items. *Journal of Educational Measurement, 36 (3), 185-198.*

Beatty, P. & Willis, G.B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly, 71* (2), 287-311.

Carmines, E. G. & Zeller, R. A. (1979). *Reliability and validity assessment.* London: Sage.

Castillo, M., Padilla, J. L., Gómez-Benito, J., & Andrés, A. (2010). A productivity map of cognitive pre-test methods for improving survey questions. *Psicothema*, 22, 482-488.

Ercikan, K.; Arim, R.; Law, D.; Domene, J. Gagnon, F. & Lacroix, S. (2010). Application of Think Aloud Protocols for Examining and Confirming Sources of Differential Item Functioning Identified by Expert Reviews. *Educational Measurement: Issues and Practice, 29 (2), 24-35.*

Ferne, T., & Rupp, A.A. (2007). A synthesis of 15 years of research on DIF in language testing: Methodo- logical advances, challenges, and recommendations. *Language Assessment Quarterly, 4,* 113-148.

Hambleton, R. K., Merenda, P., & Spielberger, C. (Eds.). (2005). *Adapting educational and psychological tests for cross-cultural assessment.* Hillsdale, NJ: Lawrence S. Erlbaum.

International Test Commission (2010). *Guidelines for translating and adapting tests*. Downloaded from the world wide web at http://www.intestcom.org on October 4, 2010.

Miller, K. (2007). Design and Analysis of Cognitive Interviews for Cross-National Testing. European Survey Research Association Annual Meeting. Prague, Czechoslovakia.

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

Millsap, R. E. & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement 17***,** 297-334.

Muthén, B. O. (1989). Using item-specific instructional information in achievement modeling. *Psychometrika*. 135-396.

Organisation for Economic Co-operation and Development. (2004). Programme for the International Assessment of Adult Competencies (PIAAC). Policy Objectives, Strategic Options and Cost Implications. Stockholm: Author.

Organisation for Economic Co-operation and Development. (2006). Literacy skills for the world of tomorrow—further results from PISA 2003. Paris: Author.

Padilla, J. L., Pérez, C., & Gonzalez, A. (1998). La explicación del sesgo en los ítems de rendimiento [The explannation of bias in achievement ítems]. *Psicothema, 10,* 481-490.

Penfield, R. D. & Algina, J. (2003). Applying the Liu-Agresti estimator of the cumulative common odds ratio to DIF detection in polytomous items. *Journal of Educational Measurement, 40,* 353-370.

Penfield, R. D. (2005). DIFAS: Differential Item Functioning Analysis System. *Applied Psychological Measurement, 29,* 150-151.

Penfield, R. D. (2007). *DIFAS 4.0 user's manual*. Downloaded from the world wide web on July 14, 2010 from http://www.education.miami.edu/facultysites/penfield/index.html.

Penfield, R. D. (2010). Distinguishing between net and global DIF in polytomous items. *Journal of Educational Measurement, 47,* 129-149.

Penfield, R. D., Gattamorta, K. & Childs, R. A. (2009), An NCME Instructional Module on Using Differential Step Functioning to Refine the Analysis of DIF in Polytomous Items. *Educational Measurement: Issues and Practice, 28,* 38–49.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of educational statistics, 4 (3),* 207-230.

Sireci, S. G. (1997). Problems and issues in linking tests across languages. *Educational Measurement: Issues and Practice, 16(1),* 12-19.

Sireci, S.G., Patsula, L., y Hambleton, R.K. (2005). Statistical methods for identifying flaws in the test adaptation process. En R.K. Hambleton, P.F. Merenda y S.D. Spielberger (eds.): *Adapting educational and psychological tests for cross-cultural assessment* (pp. 93-115). New Jersey: Lawrence Erlbaum Associates.

Schmitt, A.P. y Dorans, N.J. (1990). Differential item functioning for minority examinees on the SAT. *Journal of Educational Measurement*, 27, 67-81.

SPSS, Inc. 2007. SPSS-16 User's guide. Chicago, USA.

Willis, G. B. (2005). *Cognitive interviewing*. Thousand Oaks: Sage Publications.

Zumbo, B. D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-Type (Ordinal) Item Scores.* Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.