CLEAR Exam Review

Volume XXIV, Number 1
Spring 2014

A Journal

CLEAR Exam Review is a journal, published twice a year, reviewing issues affecting testing and credentialing. CER is published by the Council on Licensure, Enforcement, and Regulation, 403 Marquis Ave., Suite 200, Lexington, KY 40502.

Design and composition of this journal have been underwritten by Prometric, which specializes in the design, development, and full-service operation of high-quality licensing, certification and other adult examination programs.

Subscriptions to CER are sent free of charge to all CLEAR members and are available for $30 per year to others. Contact Stephanie Thompson at (859) 269-1802, or at her e-mail address, sthompson@clearhq.org, for membership and subscription information.

Advertisements and Classified (e.g., position vacancies) for CER may be reserved by contacting Janet Horne at the address or phone number noted above. Ads are limited in size to 1/4 or 1/2 page, and cost $100 or $200, respectively, per issue.

CLEAR Exam Review

VOLUME XXIV, NUMBER 1 SPRING 2014

Contents

 ISSN 1076-8025

Comparison of English and Spanish Translations of a National Certification Examination

HONG QIAN, XIAO LUO, ADA WOO, PHILIP DICKISON, AND DOYOUNG KIM
National Council of State Boards of Nursing

Abstract

The objective of this study was to investigate the extent to which the Spanish translation and socio-cultural context impacts examinee performance on a national certification examination for entry-level healthcare workers when compared to the performance of the English-speaking examinees. Thirty-three items across six forms in 2010 and 2011 were identified by three differential item functioning (DIF) procedures: Mantel-Haenszel (MH) Procedure, Rasch separate calibration t-test, and Rasch separate calibration t-test with bootstrapping. The results from this study showed that the MH Procedure and Rasch separate calibration t-test have identified similar numbers of DIF items, while the bootstrapping method identified fewer DIF items. Qualitative reviews of the 33 DIF items indicated they were not biased and the differences were meaningful to the target construct. The implications of this study are that impact, DIF and bias are not necessarily the same and that the adaptation of an examination from one language to another requires careful checking of all items before assuming the examinations are comparable.

Keywords: impact, differential item functioning, bias, Rasch separate calibration, bootstrapping

Author Note

Hong Qian, Xiao Luo, Ada Woo, Philip Dickison, and Doyoung Kim, Examinations Department, National Council of State Boards of Nursing.

The authors want to thank Marijana Dragan and Dr. Sarah Hagge for their contributions to this article.

Correspondence concerning this article should be addressed to Hong Qian, Examinations Department, National Council of State Boards of Nursing, 111 E. Wacker Drive, Suite 2900, Chicago, IL 60601-4277. E-mail: hqian@ncsbn.org

Comparison of English and Spanish Translations of a National Certification Examination

This study was based on the data from a national certification examination for entry-level healthcare workers. The examination was developed in compliance with industry standards. The test plan was based on a job analysis for entry-level healthcare workers. The item writers were selected from volunteers who applied to participate in the item development of the examination. Items were reviewed by several review panels to make sure the examination was psychometrically sound and legally defensible.

The examination is offered in English and Spanish languages. It is found that there were noticeable differences in the passing rates for English-speaking examinees and Spanish-speaking examinees as shown in Table 1.

As can be seen in Table 1, the differences are fairly large. Was this the result of unfairness? Not necessarily, because it is possible that Spanish-speaking examinees are less able than their English-speaking counterparts. If one group is less able than another group, it is not surprising that one group should perform worse than another on the examination. This is called the impact of examinations for different groups and is not necessarily due to unfairness or bias. In order to investigate the possibility of bias, we need to compare the performance of examinees from different groups having the same level of ability. If examinees of equal ability from different groups have unequal probabilities of responding to items correctly, this may indicate unfairness. This issue is usually investigated by differential item functioning (DIF) analysis.

One purpose of the study was to investigate the extent to which the Spanish translation and socio-cultural context impacts examinee performance on the examination when compared to the performance of the English-speaking examinees. More specifically, these two research questions were asked:

1. How comparable are the Spanish-translated items to the English items?

2. To what extent do candidates of the same ability level have the same level of performance on Spanish items as compared to English?

Method

In order to ensure the fairness of an examination, it is important that the two language forms are comparable in terms of item difficulties and also in terms of the overall construct that is measured. Previous research provides some support for construct equivalence across English and Spanish versions of healthcare certification exams. For example, in 2003 the comparability of scores on a national certification examination for nurse aides was examined across English and Spanish languages and two different administration condition groups for both calibration and validation samples (Wang, Wang, & Hoadley, 2003). The results showed that factor structure validities of the examination were well supported across the different groups. This study focused instead on the comparability of items via DIF analysis. In DIF analysis, an item is considered to be functioning differentially if examinees of equal ability from different groups have unequal probabilities of responding correctly to that item (Hambleton, Swaminathan, & Rogers, 1991).

Since different DIF methods can yield different results because of the different null hypotheses they test and different levels of power they have, it is recommended that more than one DIF method be used to find items that are consistently identified across the different methods (Ercikan, Arim, Law, Domene, Gagnon, & Lacroix, 2010). In this study, we used three DIF procedures to identify items that function differentially and investigate these items for potential bias.

TABLE 1. Passing Rates for Analyzed English and Spanish Examinations in 2010 and 2011

Year	Number of Analyzed English Exams	Analyzed English Exams' Average Pass Rate	Number of Analyzed Spanish Exams	Analyzed Spanish Exams' Average Pass Rate
2010	31,404	92%	223	71%
2011	25,920	88%	323	55%

Three DIF Procedures

Mantel-Haenszel (MH) procedure. The reason for using the MH procedure is that it is the most widely used and simplest DIF detection technique (Tay, Vermunt, & Wang, 2013). What's more, this procedure is not based on any specific item response model, so

TABLE 2. Sample Size for Each Form

Form	A	D	F	M	N	O	Total
Year	2010	2010	2010	2011	2011	2011	
English Sample	8,524	11,525	11,355	8,667	8,547	8,706	57,324
Spanish Sample	67	92	64	92	120	111	546

it is easy to compare the results with those from the second method, which is based on Rasch model. The MH procedure compares the differences in performance between the reference (English-speaking) and focal (Spanish-speaking) groups after they have been matched on ability (having the same total score). It is a series of analyses of two-by-two contingency tables, one for each observed score point. We used DIFAS 5.0 (Penfield, 2012) software to run the MH statistics.

Rasch separate calibration t-test. The Rasch separate calibration t-test compares the difference between difficulty estimates from two subgroups of interest for the same item. In order to assess DIF on English and Spanish forms, the Rasch separate calibration t-test was performed on six chosen forms from 2010 and 2011. Item difficulties from two groups were calibrated, and t-tests showed whether the difference was statistically significant. This method was chosen because it uses Rasch item difficulties, which were readily available for all forms of the examination.

Rasch separate calibration t-test with bootstrapping. Usually, the focal group (typically the minority group) in the DIF analysis has a small sample size (Sinharay et al., 2009). It is well known that a small sample size would result in inaccurate IRT estimates (Swaminathan & Gifford, 1983). This issue can be alleviated by incorporating Bayesian priors in estimation (Swaminathan & Gifford, 2003); however, when the prior is not evident, using arbitrary priors could introduce biases in estimation instead. In this case, the bootstrapping is more appropriate than the Bayesian approach (Diaconis & Efron, 1983; Efron, 1979; Efron & Tibshirani, 1986). Essentially, bootstrapping is used to conduct a serial resampling from the observed sample so as to construct a "sampling distribution". The bootstrapping method is a non-parametric method, which makes no assumption about the distribution of the statistic of interest as in the Bayesian approach.

In this study, the sample size of the Spanish group is small for each form. Item parameters calibrated from these samples may be too inaccurate to reveal authentic DIF effects. Therefore, for each test form, the Spanish group was bootstrapped 1,000 times. In each cycle of bootstrapping, a new sample was drawn with replacement from the Spanish test takers. The resampled responses were appended to existing observed responses of the English group to form a new data set. The new data set was analyzed in Winsteps to detect the presence of DIF. If the t-value of the DIF effect was greater than 2.58 or less than -2.58, then the item was considered as showing DIF. Examining the presence of DIF

TABLE 3. The DIF Items in 2010 Forms

	Form A			Form D			Form F		
	MH	Rasch	Rasch (bootstraping)	MH	Rasch	Rasch (bootstrapping)	MH	Rasch	Rasch (bootstraping)
Item#	5				3	3		7	7
	24	24	24		5	5		22	22
	30	30	30		9		26	26	
	35				11			27	
		37		12	12	12		33	
		49		15	15	15		37	
		55		21	21	21	40		
	59	59		25	25		42	42	42
				29	29	29	44	44	44
					30		51	51	51
				31				52	52
				34			54		
				36					
				40	40	40			
					47				
				54	54	54			
					55				
Total	5	6	2	10	14	8	6	10	6

TABLE 4. The DIF Items in 2011 Forms

	Form M			Form N			Form O		
	MH	Rasch	Rasch (bootstraping)	MH	Rasch	Rasch (bootstrapping)	MH	Rasch	Rasch (bootstraping)
Item#	1	1	1	3	3	3	15	15	15
	3	3	3	4	4	4		19	
	4	4	4		7	7	22	22	
	5	5	5	8			23		
	13	13	13		14	14		26	26
	14			17			27		
	16	16	16	21			28		
	19				23		42		
	20				25	25	44		
	22	22	22		31	31		47	47
	23	23		32	32	32	53	53	53
	26			33			60	60	
		28		35					
	29			36	36	36			
	36	36	36	38	38				
	37	37	37	40					
	40			41					
	42			43	43	43			
	44	44	44	47	47	47			
	49				48	48			
	51	51	51		50	50			
		54			52	52			
	55	55	55	53	53				
	57			54	54	54			
	58	58	58	57	57	57			
	59								
Total	24	16	13	17	18	15	9	7	4

across 1,000 bootstrapping replications, the "probability of DIF" was obtained. For example, if an item showed DIF in 800 bootstrapping replications out of 1,000, it would suggest that the presence of DIF was highly likely for that item. On the other hand, if an item showed DIF in only 200 bootstrapping replications, the DIF might be due to sampling error. This study used 700 as the arbitrary criterion to indicate the presence of DIF.

Sample

Six examination forms from 2010 and 2011, each offered in both English and Spanish, were analyzed for DIF. The test consists of 60 operational and 10 pretest items per form. Pretest items are not scored, meaning that they do not contribute to the examinee's final score and do not impact the pass or fail decision. Pretest items were not included in the DIF analysis. Each pair of forms that were compared and analyzed contained exactly the same operational items; the only difference was the language of the form. According to the current test industry standard, the examination forms are first constructed in English. As a next step, some of those English forms are then sent to a translation agency to be translated into Spanish. Those Spanish forms are later administered to the Spanish examinees. The sample consisted of data on 546 examinations given in Spanish and 57,324 in English. The sample size for each form is listed in Table 2.

Results

The DIF items identified by each DIF procedure are listed in Table 3 and Table 4.

From Table 3 and Table 4, the MH Procedure and Rasch separate calibration t-test have identified the same number of DIF items across 6 forms: 71 items in total. Compared with the conventional Rasch separate calibration t-test, the bootstrapping methods identified fewer DIF items, and more importantly, items identified by the bootstrapping methods were all identified by the conventional method. This might suggest that the conventional method tended to overestimate

TABLE 5. The Number of DIF Items identified by three procedures

Form	Number of DIF Items
A	2
D	6
F	3
M	12
N	8
O	2
Total	33

the DIF effect. The number of DIF items in each form identified by three procedures is listed in Table 5.

TABLE 6. The Direction of the DIF

Form	N	Items Favoring Spanish N	Items Favoring Spanish %	Items Favoring English N	Items Favoring English %
A	2	1	50.0	1	50.0
D	6	3	50.0	3	50.0
F	3	1	33.3	2	66.7
M	12	8	66.7	4	33.3
N	8	4	50.0	4	50.0
O	2	0	0.0	2	100.0
Overall	33	**17**	51.5	**16**	48.5

Since these items are consistently identified as DIF items by three procedures, we investigated the direction of the DIF for these items. This analysis involves exploring whether the DIF items are favoring the English-speaking group or the Spanish-speaking group. If all DIF items were consistently in favor of one group (easier for this group), this would clearly suggest a problem. As shown in Table 6, the DIF items favored each group evenly, with one more item favoring the Spanish-speaking group.

Since 33 items across 6 forms were identified as DIF items, some decisions about the future use of these items seemed warranted. Should these items be eliminated, revised, or kept as they are? Obviously, items should be eliminated or revised if they are biased. But the presence of DIF does not necessarily imply that an item is biased. A DIF item should be removed if group differences are related to construct-irrelevant features of the item. But a DIF item is not biased if the differences in performance are related to the target construct being measured. For example, an item on obstetrics may show DIF for male and female nurse aid candidates, but if the difference in performance is due to a real difference in the target construct (for example, if females have stronger knowledge in this content area), this item is not biased and should be kept as is. To determine whether a group difference is meaningful, a qualitative review of DIF items is needed. In this study, the content of the DIF items was reviewed by a staff member who speaks both English and Spanish. After reviewing and analyzing the text of the 33 items in both English and Spanish, we concluded that the content of the items with large DIF does not contain knowledge that could be inherent in one culture and different for the other. Items appeared to be measuring the same content and did not contain culture specific references.

The results of the DIF analysis suggest that this examination is not biased towards English or Spanish language test takers. Approximately 9% of the items across the six forms exhibited statistically significant DIF by three detection procedures, and these DIF items did not uniformly favor either language. The differences in performance of the few items with consistent DIF could be attributed to the real differences in the target construct.

References

Diaconis, P., & Efron, B. (1983). Computer-intensive methods in statistics. *Scientific American,* 248(5), 116-130.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics,* 1- 26.

Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science,* 54-75.

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.

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

Panfield, R. (2012). Differential item functioning analysis system (DIFAS 5.0). https://erm.uncg.edu/people/measurement-software/

Sinharay, S., Dorans, N. J., Grant, M. C., & Blew, E. O. (2009). Using past data to enhance small sample DIF estimation: A Bayesian approach. *Journal of Educational and Behavioral Statistics,* 34(1), 74-96.

Swaminathan, H., & Gifford, J. A. (1983). Estimation of parameters in the three-parameter latent trait model. In D. Weiss (Ed.), *New Horizons in Testing* (pp. 13-30). New York: Academic Press.

Swaminathan, H., Hambleton, R. K., Sireci, S. G., Xing, D., & Rizavi, S. M. (2003). Small sample estimation in dichotomous item response models: Effect of priors based on judgmental information on the accuracy of item parameter estimates. *Applied Psychological Measurement,* 27(1), 27-51.

Tay, L., Vermunt, J.K., & Wang, C. (2013). Assessing the item response theory with covariate (IRT-C) procedure for ascertaining differential item functioning. *International Journal of Testing,* 13(3), 201-222.

Wang, S., Wang, N., & Hoadley, D. (2003). Construct equivalence of a National Certification Examination that uses dual languages and audio assistant. *The Annual Meeting of the American Educational Research Association,* Chicago, IL.