

Detecting Cross-Cultural Differential Item Functioning for Increasing Validity: An Example from the American Board of Family Medicine In-Training Examination

Xian Wu, Rongxiu Wu, Michael R. Peabody¹, Thomas R. O'Neill¹

Department of Education, School, and Counseling Psychology, College of Education, University of Kentucky, ¹American Board of Family Medicine, Lexington, Kentucky, USA

Abstract

Background: The present study describes the process used to detect items for cross-cultural differential item functioning (cc-DIF) and attempts to understand cc-DIF by both statistical analysis and content review using a cultural lens. **Methods:** Data from the 2014 American Board of Family Medicine (ABFM) In-Training Examination (ITE). **Results:** cc-DIF existed in ten items on the 2014 ABFM ITE and could not be eliminated over the residency program years. International medical school graduates were benefited by seven items, whereas the United States medical school graduates (USMGs) were benefited by three items. **Discussion:** Cultural specificities and differential content familiarity likely are the primary reasons for items exhibiting cc-DIF. **Conclusions:** Investigating cc-DIF is recommended for any examination involving multicultural groups. Further, items exhibiting cc-DIF offer opportunities for students to reflect on their implicit cultural differences that may ultimately affect how they practice medicine in a multicultural society.

Keywords: Assessment, culture, family medicine, measurement, medical certification, psychometrics, research methods, testing, validity

INTRODUCTION

Test validity refers to the extent to which a test accurately measures what it is supposed to measure. Validity is not a property of the test or test scores but rather the proposed interpretation and uses of the test scores.^[1] A high-quality test should be fair across demographic groups of examinees; however, when a test shows signs of unfairly favoring one subgroup, (e.g., male vs. female), it is often assumed that the validity of those test-based inferences is threatened.^[2] Differential item functioning (DIF) exists when examinees from different groups having a differing probability or likelihood of success on an item after their levels of ability which are measured in the test have been matched.^[3-6] The Standards for Educational and Psychological Testing suggest validity should be the foremost concern of test development and use, such that evidence from many sources, including DIF analyses, is recommended for meaningful and appropriate interpretation of a test's scores.^[7]

DIF investigation refers to detecting whether test items function differently for different groups of examinees, which may be linked

systematically to the personal characteristics of the examinees and unrelated to the test's central construct.^[2] As the most popular types, gender and ethnicity DIF have been reflected in numerous studies.^[8-11] Rothman *et al.* offered some interesting examples of gender DIF in multiple-station tests of clinical skills.^[10] The authors adopted data from 23 stations used in the selection of seven successive cohorts (1987–1993) of candidates to the Ontario Preinternship Program for graduates of foreign medical schools and illustrated that an item consisting of a physical examination of a man favored the male examinees and an item consisting of an interview with a mother of an infant favored female examinees.^[10] To illustrate DIF based on ethnicity, Woo and Dragan showed how an examinee's cultural background might elicit a particular response to an item, thereby inducing DIF.^[11]

Address for correspondence: Ms. Xian Wu,

Department of Education, School, and Counseling Psychology,
College of Education, University of Kentucky, Lexington, KY, 40506, USA.

E-mail: xian.wu@uky.edu

Access this article online

Quick Response Code:



Website:
www.ehpjournal.com

DOI:
10.4103/EHP.EHP_12_18

This is an open access journal, and articles are distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 License, which allows others to remix, tweak, and build upon the work non-commercially, as long as appropriate credit is given and the new creations are licensed under the identical terms.

For reprints contact: reprints@medknow.com

How to cite this article: Wu X, Wu R, Peabody MR, O'Neill TR. Detecting cross-cultural differential item functioning for increasing validity: An example from the American board of family medicine in-training examination. *Educ Health Prof* 2018;1:19-23.

Beyond gender and ethnicity DIF, cross-cultural DIF (cc-DIF) has increasingly received attention from researchers in a variety of fields.^[8,12-16] In some international comparison studies, the terms cc-DIF and cross-national DIF are used interchangeably. Culture refers to a broader category, beyond location and ethnicity, to include any group of people who share common lifestyle characteristics that are passed on to members of the group, such as socioeconomic status and religion.^[17]

Researchers have shown that not all items exhibit measurement invariance across cultures.^[13,15,18] Measurement invariance refers to measurement operations yielding measures of the same attribute that does not vary across different conditions and situations.^[19] When an item functions differently in a specific culture in comparison with the item functioning across other cultures, cc-DIF appears.^[15] The four most likely reasons for explaining cc-DIF are (1) language translation,^[20] (2) cultural specificities,^[11,12] (3) differential curriculum coverage,^[21] and (4) differential content familiarity.^[22] DIF across cultures caused by language translation and differential curriculum coverage are generally easier to identify and correct, whereas DIF introduced by cultural specificities and differential content familiarity can be implicit and require more attention. Further, cc-DIF has been found to negatively influence both the accuracy of scores and the inferences made about them.^[8,12,15] Therefore, investigating cc-DIF can help reduce item bias and increase validity.^[2,23]

Given the importance of measurement invariance and test validity, the purpose of this study was to draw attention to the issue of cc-DIF and demonstrate how to detect cc-DIF using a real dataset from a medical certification board located in the United States (US). More specifically, we describe a process used to detect items exhibiting cc-DIF and illustrate how to discern cc-DIF by way of both a statistical and content review. Three hypotheses were formulated: (1) cc-DIF exists in a dataset taken from a diverse population of examinees, (2) cc-DIF may be persistent and could be difficult to eliminate over time, and (3) cultural specificities and differential content familiarity likely are the primary reasons for many items exhibiting cc-DIF; therefore, closer attention is required to appropriately interpret cc-DIF and reduce item bias.

METHODS

Participants

The American Board of Family Medicine (ABFM) offers residents enrolled in the Accreditation Council for Graduate Medical Education (ACGME) accredited residency programs the opportunity to take the In-Training Examination (ITE).

For the present study, two distinct participant cohorts were examined: the United States medical school graduates (USMGs) and international medical school graduates (IMGs). The sample size consisted of 10,528 participants from 3 different program years (PGYs) [Table 1]. In total, 33% of participants were IMGs, and 67% of participants were USMGs.

Instrumentation

The ITE is a low-stakes, internet-based examination that is administered annually during the last full week of October to all physician residents in ACGME accredited family medicine residency programs in the US. The examination contains 240 multiple-choice items intended to measure a physician's clinical decision-making ability as it relates to the practice of family medicine. The ITE is constructed using the same test plan specifications as the ABFMs Family Medicine Certification Examination (FMCE). Passing the FMCE is one of the requirements for receiving certification from the ABFM.^[24] The purposes of the examination are threefold: (1) to help examinees become familiar with the general format and item writing style of ABFM examinations, including the FMCE, (2) to provide an opportunity for residents to assess how well they are progressing toward ultimately passing the FMCE, and (3) to provide residency programs with comparative data about their residents' collective performance. The data utilized in this study were obtained from the 2014 ITE.

Data analysis

Examination data were scored using the dichotomous Rasch model.^[25] The Rasch separate calibration *t*-test method was employed to examine cc-DIF.^[26] Given the pairs of item calibrations and the associated estimates of standard errors, *t* statistic can be constructed using the following formula:

$$t_i = \frac{d_{i1} - d_{i2}}{\sqrt{(s_{i1}^2 + s_{i2}^2)}}$$

Where d_{i1} is the difficulty of item *i* from the calibration based on subpopulation 1, d_{i2} is the difficulty of item *i* from the calibration based on subpopulation 2, s_{i1} is the standard error of estimate for d_{i1} and s_{i2} is the standard error of estimate for d_{i2} .^[27]

In general, the criteria for the Rasch separate calibration *t*-test method involve setting the magnitude for the difference between the two calibrations ($\Delta\theta = d_{i1} - d_{i2}$) and an appropriate *p* value for significance. For instance, flagging criteria for the National Council Licensure Examination (NCLEX) were set at a $\Delta\theta \geq 0.5$ logits and $p \leq 0.0001$.^[11] In the present study,

Table 1: Descriptive summary of participants by program year

PGYs	IMGs (percentage of PGYs)	USMGs (percentage of PGYs)	Row total (percentage of total)
PGY-1	1165 (32)	2454 (68)	3619 (34)
PGY-2	1187 (35)	2356 (65)	3543 (34)
PGY-3	1160 (34)	2206 (66)	3366 (32)
Column total (percentage of total)	3512 (33)	7016 (67)	10,528

PGYs: Program years, IMGs: International medical school graduates, USMGs: United States medical school graduates

higher flagging criteria were set at a $\Delta\theta \geq 0.8$ logits and $p \leq 0.0001$ for the cc-DIF investigation. Requirements of large calibration differences and small p values are to increase the likelihood that the results are not simply due to chance. Winsteps (version 3.92.1) was utilized to estimate parameters and perform Rasch separate calibration t -tests.

Content review

Following the statistical analysis, a content review of each item flagged for DIF was conducted to determine whether the source if DIF is a threat to test validity or simply a statistical artifact. The content review was conducted by content experts who examine the items to determine whether any aspects of the item’s stem or distractor options may have had an undue influence on the responses of examinees of differing cultural backgrounds. Due to space limitations, we provide results for two items selected as examples.

RESULTS

Of the 240 items administered, 13 items (5% of total items) were flagged as candidates for cc-DIF. Eight items benefited IMGs, while five items benefited USMGs. Ten items (4% of total items) were continuously flagged over each of the 3 PGYs. Among the continuously flagged items, seven favored IMGs over USMGs [Table 2]. The items exhibiting cc-DIF appeared in multiple modules, each measuring different areas of knowledge and skills relating to family medicine.

Following the content review, items 75 and 228 were selected as examples to illustrate DIF. Item 75 provides an example of a bias against USMGs. Negative calibration differences (between -0.850 and -0.957) indicate that the IMG group was more likely to answer the question correctly than the USMG group [Table 3 and Figure 1].

Item 228 provides an example of an item bias against IMGs. Positive calibration differences (between 0.841 and 1.221)

indicate that the IMG group was more likely to answer the question incorrectly than the USMG group [Table 4 and Figure 2].

DISCUSSION

The present study describes the process used to detect items for cc-DIF and attempts to understand cc-DIF by both statistical analysis and content review using a cultural lens. Results confirm the three aforementioned hypotheses. Specifically, we found evidence of cc-DIF in the 2014 ABFMs ITE in which 33% of participants were IMGs. Among the 240 items, a total of 13 (5% of the total) were flagged, eight items benefited IMGs and five items benefited USMGs. Large calibration differences between the IMG and USMG groups ($\Delta\theta \geq 0.8$ logits) reached the high criterion for significance ($p \leq 0.0001$) indicating the findings likely are not due to chance.

cc-DIF has persistency, meaning it would be sustained and difficult to eliminate over years as shown by the ten items (4% of total items) which were continuously flagged over the 3 PGYs. These ten items did not belong to any single content category, and the cc-DIF was unlikely to be caused by lacking any specific knowledge or skill that can be made up by receiving additional training. Differences in responses caused by cultural specificities can be implicit and require more attention. IMGs, even those enrolled in the US residency programs for multiple years, may have kept certain concepts and beliefs shaped by cultural specificities. For instance, item 75 was easier for IMGs and related to the disease, goiter. Goiter was previously common in some areas that were deficient in iodine in the soil and is still prevalent in India, China, Central Asia, and Central Africa.^[28] Hence, in terms of epidemiology, IMGs might have more opportunities for exposure to goiter or learn more about goiter through curricula emphasized at their local medical schools. On the contrary, item 228 was easier for USMGs and related to exercise during pregnancy. Vigorous exercise is not a risk factor for stillbirth; however, some cultures may view it as an inappropriate and a cause of

Table 2: Flagged items for total participants and different program years

Participants	Flagged items	
	In favor of IMGs	In favor of USMGs
Total	#46, #75, #134, #142, #159, #164, #187, #232 (8 items)	#13, #52, #146, #162, #228 (5 items)
PGY-1	#46, #75, #117, #134, #142, #159, #164, #187, #232, #236 (10 items)	#7, #13, #61, #66, #144, #146, #228 (7 items)
PGY-2	#17, #44, #46, #75, #134, #142, #159, #171, #187, #232 (10 items)	#13, #52, #146, #162, #208, #215, #225, #228 (8 items)
PGY-3	#17, #44, #46, #75, #117, #134, #142, #159, #164, #187, #232 (11 items)	#13, #52, #146, #162, #225, #228 (6 items)
Continuously flagged items	#46, #75, #134, #142, #159, #187, #232 (7 items)	#13, #146, #228 (3 items)

PGYs: Program years, IMGs: International medical school graduates, USMGs: United States medical school graduates

75. A 45-year-old female had myalgias, a sore throat, and a fever 2 weeks ago. She now has anterior neck tenderness and swelling, with pain radiating up to her ears. Your examination reveals a tender goiter.

Which one of the following would support a diagnosis of subacute granulomatous thyroiditis?

A) Pretibial myxedema

B) Exophthalmos

C) Multiple nodules on ultrasonography

D) Low radioactive iodine uptake (<5%)

Figure 1: Item 75 is an example of an item flagged for content review due to bias against USMGs

Table 3: Differential item functioning information for the flagged item 75

PGYs	Focal group	Calibration (d_{11})	SE (s_{11})	Reference group	Calibration (d_{12})	SE (s_{12})	Difference ($\Delta\theta$)	JSE	<i>t</i>	<i>p</i>	Favor group
PGY-1	IMGs	-0.144	0.062	USMGs	0.707	0.041	-0.850	0.074	-11.4	<0.001	IMGs
PGY-2	IMGs	0.064	0.062	USMGs	0.976	0.042	-0.913	0.075	-12.2	<0.001	IMGs
PGY-3	IMGs	0.210	0.063	USMGs	1.167	0.043	-0.957	0.077	-12.5	<0.001	IMGs

PGYs: Program years, IMGs: International medical school graduates, USMGs: United States medical school graduates, SE: Standard error, JSE: Joint standard error

Table 4: Differential item functioning information for the flagged item 228

PGYs	Focal group	Calibration (d_{11})	SE (s_{11})	Reference group	Calibration (d_{12})	SE (s_{12})	Difference ($\Delta\theta$)	JSE	<i>t</i>	<i>p</i>	Favor group
PGY-1	IMGs	-0.302	0.064	USMGs	-1.143	0.058	0.841	0.086	9.8	<0.001	USMGs
PGY-2	IMGs	-0.030	0.063	USMGs	-1.168	0.066	1.138	0.091	12.5	<0.001	USMGs
PGY-3	IMGs	-0.001	0.066	USMGs	-1.222	0.074	1.221	0.099	12.4	<0.001	USMGs

PGYs: Program years, IMGs: International medical school graduates, USMGs: United States medical school graduates, SE: Standard error, JSE: Joint standard error

228. Which one of the following is NOT a risk factor for stillbirth?

- A) Smoking
- B) Advanced maternal age
- C) Congenital anomalies
- D) Vigorous exercise
- E) BMI > 30 kg/m²

Figure 2: Item 228 is an example of an item flagged for content review due to bias against IMGs

miscarriage and preterm birth.^[29] Similarly, Woo and Dragan also reported test items exhibiting cc-DIF. In their study, an item was found have cc-DIF because one of the response options “the client drinks warm tea before bedtime” was perceived as a correct answer by Hispanic candidates.^[11]

Culture implicitly shapes individual concepts, beliefs, motivations, experiences, and in turn behaviors. Haydel and Roeser stated that individual differences in performance on tests could be a function of a series of moment-to-moment transactions between persons and items.^[30] Therefore, examinees do not only bring their abilities into a testing situation but also other comparative characteristics shaped by differential cultural perspectives that might cause item bias. For this reason, it is important to not only investigate for gender and ethnicity DIF but also cc-DIF to reduce item bias and increase test validity. cc-DIF investigations should not be limited to international examinations. Due to the diverse population in the US, national examinations should also be candidates for cc-DIF investigations.

Although a great deal of time and expenses typically is involved in constructing a single test item, the presence of items exhibiting cc-DIF is not necessarily distressing.

Sometimes, items exhibiting DIF should be removed from the item pools. In other instances, a revision or refinement is all that is necessary to produce a psychometrically sound item.^[2] More techniques are recommended for understanding cc-DIF and revising items (e.g., the “think aloud” technique, etc.).^[11,12] In actuality, items exhibiting cc-DIF may be potentially helpful for conceptual awareness and change in medical education. In terms of the Conceptual Change Model, students replace new concepts starting from awareness and dissatisfaction with the current concepts.^[31] Items with cc-DIF offer good opportunities for educators to encourage students to reflect on their implicit beliefs and consider the reasons for their responses.

CONCLUSIONS

Cultural specificities and differential content familiarity are major reasons why items may exhibit cc-DIF. Therefore, investigating cc-DIF is essential for examinations involving multicultural groups. cc-DIF typically results in bias that may threaten test validity. Therefore, items exhibiting cc-DIF likely will need to be excluded from scoring. However, there may be value in retaining items with cc-DIF for educational purposes. Items exhibiting cc-DIF offer opportunities for educators to intervene and encourage students to reflect on their implicit beliefs that may impact their medical practice and patient outcomes.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

REFERENCES

1. Kane M. Validation strategies: Delineating and validating proposed interpretation. In: Lane S, Raymond MR, Haladyna TM, editors. Handbook of Test Development. 2nd ed. New York: Routledge; 2016. p. 64-80.
2. Osterlind SJ, Everson HT. Differential Item Functioning. 2nd ed.

- Thousand Oaks, Calif: SAGE; 2009.
3. Clauser BE, Mazor KM. Using statistical procedures to identify differentially functioning test items. *Educ Meas* 1998;17:31-44.
4. Dorans NJ, Holland PW. DIF detection and description. In: Holland PW, Wainer H, editors. *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum; 1993. p. 35-66.
5. Millsap RE, Everson HT. Methodology review: Statistical approaches for assessing measurement bias. *Appl Meas Educ* 1993;17:297-334.
6. Wyse AE, Mapuranga R. Differential item functioning analysis using Rasch item information functions. *Int J Test* 2009;9:333-57.
7. American Educational Research Association, American Psychology Association, National Council on Measurement in Education. *Standards for Educational and Psychological Testing*. Washington DC: American Educational Research Association 2014.
8. Le LT. Investigating gender differential item functioning across countries and test languages for PISA science items. *Int J Test* 2009;9:122-33.
9. Norcini JJ, Fletcher SW, Quimby BB, Shea JA. Performance of women candidates on the American board of internal medicine certifying examination, 1973-1982. *Ann Intern Med* 1985;102:115-8.
10. Rothman AI, Cohen R, Ross J, Poldre P, Dawson B. Station gender bias in a multiple-station test of clinical skills. *Acad Med* 1995;70:42-6.
11. Woo A, Dragan M. Ensuring validity of NCLEX® with differential item functioning analysis. *J Nurs Regul* 2012;2:29-31.
12. Benitez Baena I, Padilla JL, Hidalgo Montesinos MD, Sireci SG. Using mixed methods to interpret differential item functioning. *Appl Meas Educ* 2016;29:1-16.
13. De Jong Martijn G, Steenkamp J, Benedict EM, Fox JP. Relaxing measurement invariance in cross-national consumer research using a hierarchical IRT model. *J Consum Res* 2007;34:260-78.
14. Lee H, Geisinger KF. The matching criterion purification for differential item functioning analyses in a large-scale assessment. *Educ Psychol Meas* 2016;76:141-63.
15. Sachse KA, Roppelt A, Haag N. A comparison of linking methods for estimating national trends in international comparative large-scale assessments in the presence of cross-national DIF. *J Educ Meas* 2016;53:152-71.
16. Yildirim HH, Berberoğlu G. Judgmental and statistical DIF analyses of the PISA-2003 mathematics literacy items. *Int J Test* 2009;9:108-21.
17. Rasmussen HN, Lavish L. Broad definitions of culture in the field of multicultural psychology. In: Pedrotti JT, Edwards L, editors. *Perspectives on the Intersection of Multiculturalism and Positive Psychology*. New York: Springer Science + Business Media; 2014. p. 17-30.
18. Kankaraš M, Moors G. Analysis of cross-cultural comparability of PISA 2009 scores. *J Cross Cult Psychol* 2014;45:381-99.
19. Horn JL, McArdle JJ. A practical and theoretical guide to measurement invariance in aging research. *Exp Aging Res* 1992;18:117-44.
20. Ercikan K. Translation DIF on TIMSS. Paper presented at: The Annual Meeting of the National Council on Measurement in Education. Montreal, Quebec, Canada; 19-23 April, 1999.
21. Klieme E, Baumert J. Identifying national cultures of mathematics education: Analysis of cognitive demands and differential item functioning in TIMSS. *J Educ Dev* 2001;16:385-402.
22. van de Vijver F, Tanzer NK. Bias and equivalence in cross-cultural assessment: An overview. *Eur Rev Appl Psychol* 2004;54:119-35.
23. Hambleton RK, Kanjee A. Increasing the validity of cross-cultural assessments: Use of improved methods for test adaptations. *Eur J Psychol Assess* 1995;11:147-57.
24. O'Neill TR, Li Z, Peabody MR, Lybarger M, Royal K, Puffer JC, *et al.* The predictive validity of the ABFM's in-training examination. *Fam Med* 2015;47:349-56.
25. Rasch G. *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen, Denmark: Danish Institute for Educational Research; 1960.
26. Wright BD, Stone MH. *Best Test Design: Rasch Measurement*. Chicago, IL: MESA Press; 1979.
27. Smith RM. A comparison of the Rasch separate calibration and between-fit methods of detecting item bias. *Educ Psychol Meas* 1996;56:403-18.
28. McNeil DG Jr. In raising the World's I. Q. the secret's in the salt. *New York Times*; 16 December, 2006. Available from: <http://www.nytimes.com>. [Last accessed on 2017 Jul 24].
29. The Beijing Times. Pregnant women should not do exercise causing sweating. *Sport and Health*. 10 January, 2011. Available from: <http://www.tjzx.zryhy.com.cn/Html/News/Articles/222374.html>. [Last accessed on 2017 Jul 22].
30. Haydel AM, Roeser RW. On motivation, ability, and the perceived situation in science test performance: A person-centered approach with high school students. *Educ Assess* 2002;8:163-89.
31. Posner GJ, Strike KA, Hewson PW, Gertzog WA. Accommodation of a scientific conception: Toward a theory of conceptual change. *Sci Educ* 1982;66:211-27.