**International Journal of Language Studies**
Volume 12, Number 3, July 2018, pp. 1-14

# On the impact of differential item functioning on test fairness: A Rasch modeling approach

Hossein KARAMI, University of Tehran, Iran

With the rising concerns over the fairness of psychometric tests, Differential Item Functioning (DIF) is increasingly applied in test development and use situations. Despite its widespread application, many questions remain unresolved as to the impact of DIF on the overall performance of test takers on the test and the fairness of the test. Rasch model and the invariance principle provide versatile tools for investigating the issue. This study aimed to shed light on the issue using the framework of the Rasch model. The participants were 1651 examinees (*N*=1651) who had taken a high-stakes test in 2010. A DIF analysis of the data showed that a large number of items were in fact functioning differentially for examinees from different academic backgrounds. Despite the existence of so many DIF items, the ability estimates obtained from the original test and those obtained from an item composite containing only neutral items showed that the estimates were highly invariant. The implications for validity and fairness issues are discussed in the light of the results of the study.

**Keywords:** Bias; Differential Item Functioning; Fairness; Impact of DIF; Item Analysis; Rasch; Validity

## 1. Introduction

It has become a truism to say that validity is the single most important consideration in test development and use (Bachman, 1990). Therefore, it is incumbent upon the test developers and users alike to provide evidence that their tests are appropriate for the intended purposes. In addition to the evidence required to support the inferences, such validity arguments must present a firm consideration of the counterarguments or alternative hypotheses (Mislevy, Steinberg, & Almond, 2003). These counterarguments are other ways of explaining the observed data. Therefore, if the test users intend to argue in favor of a specific inference model, then they should also attempt to show that alternative arguments are not plausible.

Messick (1989) famously argued that there are two sources of threat against validity: (1) construct underrepresentation, and (2) construct-irrelevant variance. The former happens when the totality of the construct is not

ISSN: 2157-4898; EISSN: 2157-4901

represented in the test. The latter occurs whenever factors other than the construct itself are affecting the performance on the test. Such construct-irrelevant variance will render the test biased (Salmani Nodoushan, 2008). "An item is said to be biased when test takers from one group are less likely to answer an item correctly than test takers of another group due to some characteristic of the item or the test situation that is not relevant to the purpose of the test" (Wiberg, 2007, pp. 1-2).

Construct-irrelevant variance has been investigated as part of test fairness (Karami, 2013a). In fact, a test may be said to be biased, or not fair, when part of the scores is due to construct-irrelevant factors. A variety of techniques have been developed and applied for examining test fairness. These range from various modifications of Generalizability Theory (e.g., Alavi & Karami, 2017; Karami, 2012b, 2013a,b; Salmani Nodoushan, 2009) to Diagnostic Classification Modeling (DCMs) (e.g., Rupp, Templin, & Henson, 2010), various models of Item Response Theory (e.g., De Ayala, 2009; Embretson & Reise, 2000), and numerous techniques based on Structural Equation Modeling (SEM) for assessing parameter invariance (e.g., Kline, 2015).

An overall area of research specifically focused on fairness at the item level is Differential Item Functioning (Holland & Wainer, 1993; Osterlind & Everson, 2009). DIF occurs when examinees with the same ability level but from two different groups have different probabilities of endorsing an item (Clauser & Mazor, 1998). Therefore, DIF is a necessary but not sufficient condition for bias. If the factor causing DIF is part of the construct itself, then it is called impact (Karami, 2012a). On the other hand, if the source of DIF is not part of the construct, then the item would be biased.

There have been debates over the impact of DIF on test fairness (Karami & Salmani Nodoushan, 2011). While a number of researchers (e.g., Roznowski & Reith; 1999; Zumbo, 2003) have reported little impact, others (e.g., Pae & Park, 2006) have stated that the effect may be significant. The purpose of the present study, therefore, is to investigate the impact of DIF on test validity in the context of the Rasch model and the invariance principle. In the next section, a brief overview of the Rasch model is offered. The invariance principle and its versatility for the purpose of the present study will follow.

## 2. Background

### 2.1. The Rasch model

The Rasch model is named after its developer, the Danish mathematician Georg Rasch. It provides a strong framework for the estimation of person and item parameters. It focuses on the probability of endorsing item $i$ by person $m$. In aiming to model this probability, it essentially takes into account person

ability and item difficulty. Probability is a function of the *difference* between person ability and item difficulty. The following formula shows just this:

$$\mathrm{P}\,(\mathrm{x} = 1 \mid \theta, \delta) = f(\theta_n - \delta_i)$$

where $\theta_n$ is person ability and $\delta_i$ is item difficulty. The formula simply states that the probability of endorsing the item is a function of the *difference* between person ability, $\theta_n$, and item difficulty, $\delta_i$. This is possible becasue item difficulty and person ability are on the same scale in the Rasch model. It is also intuitively appealing to conceive of probability in such terms. The Rasch model assumes that any person taking the test has an amount of the construct gauged by the test and that any item also shows an amount of the construct. These values work in the opposite direction. Thus, it is the *difference* between item difficulty and person ability that counts (Rasch, 1980).

Three cases can be considered for any encounter of persons and items (Wilson, 2005):

a. item difficulty and person ability are the same, $\theta_n - \delta_i = 0$, and the person has an equal probability of endorsing the item or failing. Thus, probability is .5.
b. person ability is greater than item difficulty, $\theta_n - \delta_i > 0$, and the person has more than .5 probability of endorsing the item.
c. person ability is lower than item difficulty, $\theta_n - \delta_i < 0$, and the probability of giving a correct response to the item is less than .5.

The exact formula for the Rasch model is the following:

$$\mathrm{Ln}\left(\frac{\mathrm{P_{ni}}}{1 - \mathrm{P_{ni}}}\right) = \theta_n - \delta_i$$

It is clear from the above formula that the probability of endorsing an item is a function of the difference between person ability and item difficulty.

### 2.2. Measurement invariance

In everyday practice, when we use a scale of measurement, the assumption is that the scale is invariant with regard to the context where it is applied and the object of measurement. For example, if we use a thermometer to gauge the temperature of the room, the expectation is that we will come up with the same measure regardless of the specific thermometers we have used. That is, the measurement of the temperature should be independent of the particular instrument used. By extension, we expect the thermometers to be invariant with respect to the context where they are utilized. For example, the

thermometer used in Iran should give the same measure as it gives in the United States—assuming, of course, that the real temperature is the same. Also, the object of measurement should not change the properties of the instrument used to gauge it. For example, the units of the meter used to measure the length of a room should not depend on which room is being measured. This outline of measurement invariance is so ubiquitous in the physical sciences that even expressing it may seem rather strange (Bond & Fox, 2007). However, the situation is totally different when it comes to the social sciences.

The dominant measurement approach in the social sciences during the early and mid-twentieth century was the Classical Test Theory. CTT is still widely applied in many places across the globe due to its simplicity (Salmani Nodoushan, 2009). However, the problem with this approach to measurement is that the results of the analysis are not sample independent (Embretson & Reise, 2000). If we give a test to a group of highly proficient learners, item facility indices will be very high indicating that the test is extremely easy. On the other hand, if the same test is given to a group of low proficiency students, the item facility indices will be low implying that the test is extremely difficult. Thus, the relative difficulty of the test is dependent on the sample to which the test is administered. That is, item parameters in CTT are sample-dependent.

The above discussion on the sample-dependence of item parameters in CTT also holds true in measuring person abilities. If a group of students are given an easy test, they will get most of the items right. Thus, they will be judged as highly proficient learners. On the other hand, if a difficult test is administered, their estimated ability levels will be much lower implying that they are not proficient enough. The ability levels of these learners have not changed but each time they take a test, they come up with different estimates due to the test-dependence of person abilities in CTT.

Contrary to CTT, the Rasch model can fulfill the promises of test-independent person ability estimates and sample-independent item parameters. Rasch called this 'specific objectivity'. Specific objectivity refers to the fact that "comparisons between objects must be generalizable to beyond the specific conditions under which they are observed" (Embretson & Reise, 2000, p. 143). Here the term 'specific' has a special sense—namely, that measures of person abilities are independent of the specific items administered, and that item parameter estimates are invariant with regard to the specific sample used to calibrate them.

Rasch, of course, was not the first to discover or even to point out such a requirement for a justified measurement scale. Thurstone (as cited in Wilson, 2005) had voiced the same concern much earlier: "The scale must transcend

the group measured. A measuring instrument must not be seriously affected in its measuring function by the object of measurement" (p. 547). However, the ideas of Thurstone were not fulfilled until Rasch proposed his famous model in the 1960s.

The forgoing discussion has shown that the estimated person abilities in the Rasch model are invariant with respect to the specific test administered. The implication is that the difference between the ability levels of any two persons is the same regardless of the specific test they take. Let's see how it is possible in the Rasch model.

As indicated earlier, the formula for the Rasch model is as follows:

$$\mathrm{Ln}\left(\frac{\mathrm{P_{ni}}}{1-\mathrm{P_{ni}}}\right)=\theta_n-\delta_i$$

The formula shows the probability of a correct response to item $i$ by person $n$. Now, the probability of correct response for person $m$ on the same item is:

$$\mathrm{Ln}\left(\frac{\mathrm{P_{mi}}}{1-\mathrm{P_{mi}}}\right)=\theta_m-\delta_i$$

Subtracting the probability of person $n$ from person $m$ yields:

$$\mathrm{Ln}\left(\frac{\mathrm{P_{mi}}}{1-\mathrm{P_{mi}}}\right)-\mathrm{Ln}\left(\frac{\mathrm{P_{ni}}}{1-\mathrm{P_{ni}}}\right)=\theta_m-\delta_i-(\theta_n-\delta_i)$$
$$=\theta_m-\theta_n$$

As is evident from the above formula, the item parameters cancel each other out, and only person ability estimates remain in estimating the difference between the ability levels of any two persons. The same argument can be put forward to show the sample-independent estimation of item parameters.

The invariance principle provides a versatile tool for investigating the impact of the DIF on test fairness. In this context, DIF can be defined as the failure of the invariance principle at the item level (Engelhard, 2009). That is, an item is flagged as showing DIF whenever it is not functioning in the same way for two groups. The impact on fairness of the inclusion of a large number of DIF items in the test can be readily investigated by examining the invariance principle at the test level. Therefore, the question of the impact of DIF on test fairness can be changed into this question: Does the item-level failure of the invariance principle lead to the failure of the invariance principle at the test level? If it does, then fairness may be undermined. Otherwise, fairness may not be under question.

As such, the present study is an attempt at investigating the impact of DIF on test fairness in the context of the invariance principle. It follows the general framework outlined above.

## 3. Method

### 3.1. Participants

The participants of the present study (*N*=1651) were sampled from a population of over 4000 examinees who had taken the University of Tehran English Proficiency Test (UTEPT) in 2010. They were coming from two different academic backgrounds: Agriculture (*N*=783) and Engineering (*N*=868). We did not have access to either the age or gender of the examinees.

### 3.2. Instrumentation

The applicants to the PhD courses of the University of Tehran are required to provide the authorities with their scores on a proficiency test called the University of Tehran English Proficiency Test (UTEPT). As a regulation, the candidates will not be allowed to sit for any PhD Entrance Exam unless they present the criterion score. Thus, a passing score on this proficiency test is a prerequisite for acceptance into any PhD program at the University of Tehran. Taking into account such serious consequences for the test takers, it is clear that the examinees were highly motivated to do their best on this test.

The UTEPT test comprises three sections including Grammar, Reading, and Vocabulary. The number of questions for each section in the 2010 version used in this study was as follows:

1. Structure and Written Expression (30 items)
2. Vocabulary (35 items)
3. Reading Comprehension (35 items)

All questions were in the multiple choice format. The Reading section comprises passages immediately followed by a number of comprehension questions. The number of comprehension questions is different for each passage. Usually, a total raw score is reported to the candidates which is simply the sum of the scores they get on the three subtests.

### 3.3. Procedure

The following steps were followed in this study:

1. Calibration of the data using the Rasch model, and checking model-data fit;

2. checking the assumptions of the Rasch model (i.e., local independence and unidimensionality);
3. DIF analysis; and
4. examining the impact of DIF on test validity.

The first two steps are an integral part of all studies where the Rasch model is used for data calibration. That is, whenever the Rasch model is applied, it is essential on the part of the researchers to check for data-model fit.

## 4. Results and discussion

### 4.1. Data calibration and fit analysis

The data were calibrated using the *Winsteps*® software, version 3.70.1.1 (Linacre, 2010b). The results indicated that all but two items fit the predictions of the Rasch model well. The outfit mean squares statistics for these two items were 1.33 and 1.66. The usual levels of acceptable fit for both statistics are between 0.7 and 1.3 (Bond & Fox, 2007). The graphs of the empirical Item Characteristic Curves (ICCs) were also checked for the two items that did not fit the model. The ICCs indicated a large degree of departure from the pattern predicted by the Rasch model. Therefore, these two items were discarded from further analysis.

### 4.2. Checking the assumptions

The unidimensionality assumption requires that any measurement scale gauge one, and only one, ability or attribute at a time. As we noted above, this assumption can never be met in its sense. As Linacre (2010a) puts it, the question is not “are my data perfectly unidimensional”—because they aren't. The question becomes “Is the lack of unidimensionality in my data sufficiently large to threaten the validity of my results?" (p. 330). Thus, we are not looking for complete unidimensionality in the data. “What is required for the unidimensionality assumption to be met adequately by a set of test data is the presence of a ‘dominant’ component or factor that influences test performance” (Hambleton, Swaminathan, & Rogers, 1991, p. 9).

*Winsteps*® runs a Principle Components Analysis (PCA) on the data to see whether one dominant factor can account for the pattern of responses. The PCA divides the total variance into two components: (1) the Rasch modeled variance, and (2) the error variance. A comparison of the amount of variance explained by the Rasch model and that explained by the first contrast will show whether the data are unidimensional.

The calibration of the data in the present study showed that the Rasch-modeled dimension accounted for 25 eigenvalues while the secondary

dimension amounted to less than 3 eigenvalues. A comparison of the disattenuated correlations among the subtests, and also the subtests and the total test, showed that they were all equal to one. Therefore, it was concluded they were "telling the same story" (Linacre, 2010b, p. 436).

Another assumption closely related to unidimensionality is local independence. This assumption requires that the response to an item should be independent of the responses to all other items. In other words, the response to an item should not affect performance on other items.

Local independence was also ensured by analyzing the standardized residual correlations reported in *Winsteps*®. The largest correlation was between two Reading items, and that amounted to a correlation of only .21 indicating that the two items shared only 4 percent of their variance. Thus, local independence holds in the data.

### 4.3. DIF Analysis

In the Rasch model, the amount of DIF is calculated by a separate calibration *t*-test approach first proposed by Wright and Stone (1979; see Smith, 2004). The formula is the following:

$$t = \frac{d_{i2} - d_{i1}}{\sqrt{({s^2}_{i2} - {s^2}_{i1})}}$$

where $d_{i1}$ is the difficulty of item *i* in calibration 1, $d_{i2}$ is the difficulty of item *i* in calibration based on group 2, ${s^2}_{i1}$ is the standard error of estimate for $d_{i1}$, and ${s^2}_{i2}$ is the standard error of estimate for $d_{i2}$.

The results of DIF analysis showed that 33 items displayed significant DIF at $p$<0.01 level. Out of these items, 17 items were favoring the Agriculture students while the rest (16 items) were in favor of the Engineering group. Therefore, we have a test with a large number of items showing DIF in favor of one of the groups. The next step in the study was to investigate the impact of so many DIF items on test validity.

### 4.4. Impact of DIF on fairness

The results of the DIF analysis indicate that there are a large number of DIF items in the test (33 percent). In other words, there are 33 instances of the failure of the invariance principle at the item level. With so many DIF items, the test users should be concerned with the fairness of the test. In order to investigate the impact of so many DIF items on test fairness—or in the context of the present study, the impact of many instances of item-level failure of the invariance principle on the test-level invariance—the ability estimates of the

examinees on two tests were compared: (a) the original test, and (b) a test comprising only neutral items. The comparability of these two sets of ability estimates would show if the inclusion of a large number of DIF items did or did not jeopardize test fairness.

The correlation coefficient between these two sets of ability estimates amounted to a value of about 0.98, indicating that the inclusion of a large number of DIF items in the UTEPT test had not jeopardize test fairness. It is certainly a high level of association between the two tests. A graphical representation of the degree of relationship between the two sets of ability estimates is displayed in Figure 1.

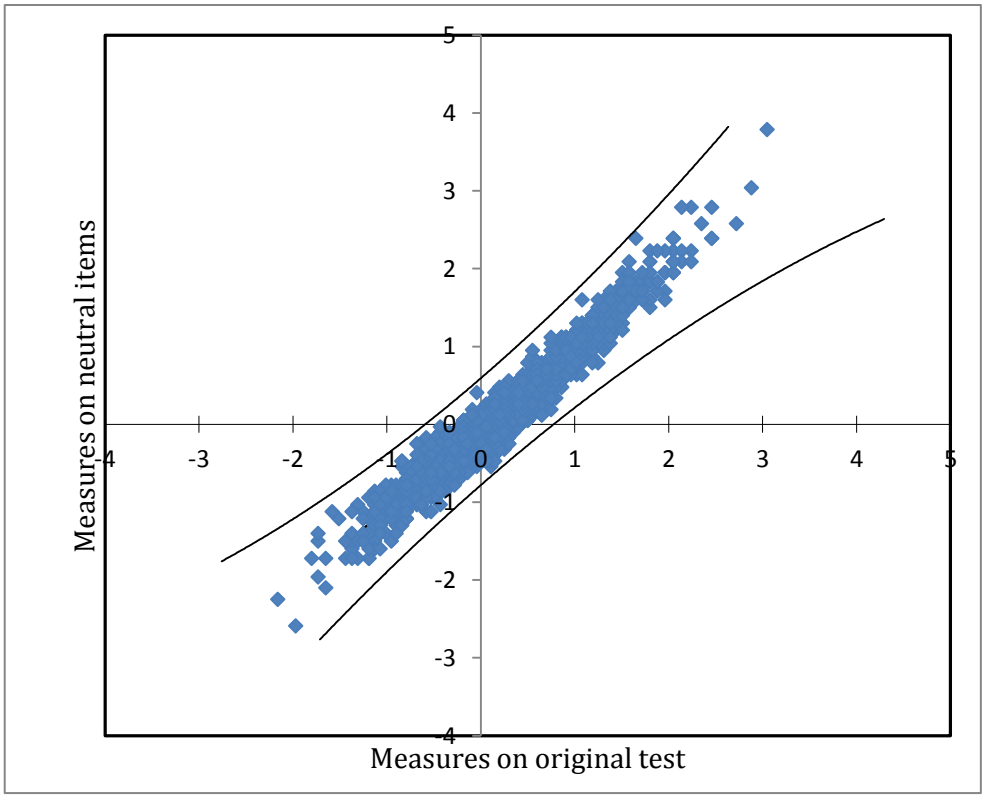


*Figure 1*. Cross-plot of the ability estimates on two tests.

Figure 1 shows a cross-plot of the ability estimates on the two tests. The sidelines show the confidence interval. If the ability estimates are invariant on the two tests, then the majority of the estimates should fall within the sidelines. In this case, virtually all the ability estimates occur within the confidence interval showing that the ability estimates are highly invariant.

In sum, it is clear that the total test is not rendered biased by the inclusion of a large number of DIF items. In other words, the invariance principle holds at the test level although there are many instances of its failure at the item level.

A number of previous research studies (e.g., Roznowski & Reith, 1999; Zumbo, 2003) have reported similar results. They have conducted similar analysis using different methodologies and have come up with similar results. Zumbo (2003), for example, used the methods of Structural Equation Modeling and confirmatory factor analysis to investigate the issue. These methods are highly complex and not understandable by all working in language testing research. The Rasch model, however, provides a much easier, albeit vigorous, approach towards investigating the issue.

However, a caveat is in order. Fairness is a multidimensional concept as previous research has indicated (Camilli, 2006; Davies, 2010; Kane, 2010; Karami, 2016; Kunnan, 2010; Xi, 2010). The fact that the inclusion of a large number of DIF items did not impact the overall performance of the examinees in this test does not necessarily behave in the same way in all other tests. Such analyses should be an integral part of all test development and use. In item banking contexts, particularly, all items should have a record of their DIF index so that in future test use situations all such items are not packed into a test.

## 5. Conclusion

The Rasch model and the invariance principle provide versatile tools for investigating the impact of DIF on test validity. Specifically, the problem can be turned into an examination of the effect of the item-level failure of the invariance principle on the test-level invariance of the ability estimates.

The findings of the present study indicated that there were 33 items displaying significant DIF in UTEPT. In other words, there were 33 instances of the failure of the invariance principle at the item level. Despite the inclusion of so many DIF items in the test, a comparison of two sets of ability estimates obtained from the original test and an item composite comprising only neutral items indicated that the ability estimates were highly invariant. It may therefore be concluded that the ability estimates based on the total test may remain invariant despite their instability at the item level. If this happens, then the validity of test is not undermined either.

There are several limitations to the present study. First, the discussion has been limited to uniform DIF where an item favors members of the reference group all along the ability scale. A comprehensive examination of the impact of DIF on test fairness should also focus on non-uniform DIF, the case where an item performs in favor of a given group up to a certain level on the ability

scale and then the relationship is reversed. Next, the data utilized here have been limited to dichotomous data. An examination of polytomously scored items may lead to different results. Furthermore, only one method was used for DIF detection. This may lead to type 1 error in flagging items as DIF. An examination of the existence of DIF by two or more DIF detection techniques may be more justified to ensure that the items are really functioning differentially for the groups.

## The Author

Hossein Karami (Email: hkarami@ut.ac.ir) is Assistant Professor of Applied Linguistics/TESOL at the English Department of the University of Tehran, Iran. His areas of interest include validity and fairness, especially in the context of language testing. His research has been published in various international scholarly journals including *Language Testing*, *International Journal of Bilingual Education and Bilingualism*, *Educational Research and Evaluation*, *RELC Journal*, *Psychological Test and Assessment Modeling*, *TESOL Journal*, *Asia-Pacific Education Review*, and *International Journal of Language Studies*.

## References

Alavi, S. M., & Karami, H. (2017). The impact of background knowledge on test performance: A multivariate G-theory approach. *International Journal of Language Studies, 11*(1), 23-44.

Amirian, S. M. R., Alavi, S. M., & Fidalgo, A. M. (2014). Detecting gender DIF with an English proficiency test in EFL context. *Iranian Journal of Language Testing, 4*(2), 187-203.

Aryadoust, V., Goh, C. C., & Kim, L. O. (2011). An investigation of differential item functioning in the MELAB listening test. *Language Assessment Quarterly, 8*(4), 361-385.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum.

Camilli, G. (2006). Test fairness. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 221-256). New York: American Council on Education & Praeger series on higher education.

Chen, Z., & Henning, G. (1985) Linguistic and cultural bias in language proficiency tests. *Language Testing, 2*(2), 155-163.

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

Davies, A. (2010). Test fairness: A response. *Language Testing, 27*(2), 171-176.

De Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: Guilford Press.

Elder, C. (1996). The effect of language background on "foreign" language test performance: The case of Chinese, Italian, and Modern Greek. *Language Learning, 46*, 233-282.

Embretson, S. E., & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum Publishers.

Engelhard, G. (2009). Using item response theory and model-data fit to conceptualize differential item and person functioning for students with disabilities. *Educational and Psychological Measurement, 69*, 4, 585-602.

Geranpayeh, A., & Kunnan, A. J. (2007) Differential item functioning in terms of age in the certificate in advanced English examination. *Language Assessment Quarterly, 4*, 190-222.

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

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Kane, M. (2010). Validity and fairness. *Language Testing, 27*, 177-182.

Karami, H. (2012a). An introduction to differential item functioning. *International Journal of Educational and Psychological Assessment, 11*(2), 59-76.

Karami, H. (2012b). The relative impact of persons, items, subtests, and academic background on performance on a high-stakes language proficiency test. *Psychological Test and Assessment Modeling, 54*(3), 211-226.

Karami, H. (2013a). An investigation of the gender differential performance on a high stakes test in Iran. *Asia Pacific Education Review, 14*(3), 435-444.

Karami, H. (2013b). The quest for fairness in language testing. *Educational Research and Evaluation, 19*(2&3), 158-169.

Karami, H. (Ed.). (2016). *Fairness issues in educational assessment*. New York: Routledge.

Karami, H., & Salmani Nodoushan, M. A. (2011). Differential item functioning (DIF): Current problems and future directions. *International Journal of Language Studies, 5*(3), 133-142.

Kim, M. (2001). Detecting DIF across the different language groups in a speaking test. *Language Testing, 18*(1), 89-114.

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). New York: Guilford Press.

Kunnan, A. J. (2010). Test fairness and Toulmin's argument structure. *Language Testing, 27*(2), 183-189.

Linacre, J. M. (2010a). *A user's guide to Winsteps*®. Retrieved July 7, 2010 from http://www.winsteps.com/.

Linacre, J. M. (2010b). *Winsteps*® (Version 3.70.0) [Computer Software]. Beaverton, Oregon: Winsteps.com.

McNamara, T., & Ryan, K. (2011). Fairness versus justice in language testing: The place of English literacy in the Australian citizenship test. *Language Assessment Quarterly, 8*(2), 161-178.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education & Macmillan.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3-62.

Osterlind, S. J., & Everson, H. T. (2009). Differential item functioning (2nd ed.). Thousand Oaks, CA: Sage.

Pae T., & Park G. P. (2006). Examining the relationship between differential item functioning and differential test functioning. *Language Testing, 23*(4), 475-496.

Pae, T. (2012). Causes of gender DIF on an EFL language test: A multiple-data analysis over nine years. *Language Testing, 29*(4), 533-554.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: The University of Chicago Press.

Roznowski, M., & Reith, J. (1999). Examining the measurement quality of tests containing differentially functioning items: do biased items result in poor measurement? *Educational and psychological Measurement, 59*(2), 248-70.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods and applications*. New York, NY: Guilford Press.

Salmani Nodoushan, M. A. (2008). Performance assessment in language testing. *Journal on School Educational Technology, 3*(4), 1-7.

Salmani Nodoushan, M. A. (2009). Measurement theory in language testing: Past traditions and current trends. *Journal on Educational Psychology, 3*(2), 1-12.

Smith, R. (2004). Detecting item bias with the Rasch model. *Journal of Applied Measurement, 5*(4), 430-449.

Wiberg, M. (2007). Measuring and detecting differential item functioning in criterion-referenced licensing test: A theoretic comparison of methods. *Educational Measurement* [technical report No. 2].

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. London: Lawrence Erlbaum Associates.

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

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing, 27*(2), 147-170.

Zumbo, B. (2003). Does item-level DIF manifest itself in scale-level analyses? Implications for translating language tests? *Language Testing, 20*, 136–47.