Differential Item Functioning (DIF): Current problems and future directions

Hossein Karami, University of Tehran, Iran
Mohammad Ali Salmani Nodoushan, IECF, Iran

With the rising concerns over the fairness of language tests, Differential Item Functioning (DIF) has been increasingly applied in bias analysis. Despite its widespread use in psychometric circles, DIF is facing a number of serious problems. This paper is an attempt to shed light on a number of the issues involved in DIF analysis. Specifically, the paper is focused on four problems: (a) the inter-method indeterminacy, (b) the intra-method indeterminacy, (d) the ad hoc interpretations, and (d) the impact of DIF on validity. In order to orient the reader, the paper also provides a brief introduction to the fundamental concepts in DIF analysis.

Keywords: Bias; Indeterminacy; Ad Hoc Interpretations; Fairness; Validity; DIF

1. Introduction

Differential Item Functioning (DIF) occurs whenever people of the same ability level but from different groups have differential probabilities of endorsing an item (Kamata & Vaughn, 2004). If the factor bringing about such a difference is not part of the construct of focus in the test, then the test would be biased. If, on the other hand, the differential performance of two groups can be attributed to a true difference in their ability levels, it is called impact rather than bias (Kamata & Vaughn, 2004).

With the rising concerns over the fairness of language tests, DIF has been increasingly applied in bias analysis. In fact, Zumbo (1999) states that DIF has become "the new standard in psychometric bias analysis" (p. 6). A plethora of research studies has applied DIF analysis to investigate the existence of bias in their tests. These studies have focused on such factors as gender (e.g., Ryan & Bachman, 1992; Karami, 2011; Shabani, 2008; Takala & Kaftandjieva, 2000), language background (Chen & Henning, 1985; Brown, 1999; Elder, 1996; Kim 2001; Ryan & Bachman, 1992), age (Geranpayeh & Kunnan, 2007) and academic background or content knowledge (Alderson & Urquhart, 1985; Hale, 1988; Karami, 2010; Pae, 2004).

Despite its widespread application in psychometric circles, DIF is facing a number of challenges in its current state. This paper will discuss a number of the most important problems in DIF analysis and will suggest some possible directions for future research. Before embarking on a discussion of these problems, however, a review of DIF analysis will be presented. This is intended to orient the reader towards the issues involved.

2. Background

As stated earlier, DIF happens whenever two groups of equal ability levels have different probabilities of correctly answering an item because they are from different groups. Therefore, DIF is a *prima facie* evidence that the possibility that the test is biased exists. That is, the existence of DIF does not necessarily mean that the test is biased. In fact, DIF is a necessary but not sufficient condition for bias (McNamara & Roever, 2006). Bias is ensured if, and only if, the source of DIF is not part of the construct of focus in the test.

Whenever an item is flagged as displaying DIF, the source of DIF should be investigated to see if it is biased or not. Any item flagged as showing DIF is biased if, and ONLY IF, the source of variance is irrelevant to the construct being measured by the test (i.e., DIF is due to construct-irrelevant variance). In other words, it is a case of construct-irrelevant variance where the groups of test takers perform differentially on an item, not because of an actual ability difference, but because of the unwanted effect of say a grouping factor (Messick, 1989, 1994).

There are at least two groups, i.e. focal and reference groups, in any DIF study. The focal group, a group of minorities, for example, is the potentially disadvantaged group. The group which is considered to be potentially advantaged by the test is called the reference group (McNamara & Roever, 2006). Note, however, that naming the groups is not always clear-cut. That is, the labeling of the groups can be arbitrary (Bachman, 2004). Moreover, there are two types of DIF, namely uniform and non-uniform DIF. Uniform DIF occurs when a group performs better than another group on all ability levels. That is, almost all members of a group outperform almost all members of the other group who are at the same ability levels. In the case of non-uniform DIF, members of one group are favored up to a level on the ability scale and, from that point on, the relationship is reversed. That is, there is an interaction between grouping and ability level (Bachman, 2004).

As stated earlier, DIF occurs when two groups *of the same ability levels* have different chances of endorsing an item. Thus, a criterion is needed for matching the examinees for ability. The matching process is called conditioning and the criterion dubbed as the matching criterion. Matching is of two types: internal and external. In the case of internal matching, the criterion is the observed or latent score of the test itself. For external matching, the observed or latent score of another test is considered as the criterion. External matching can become problematic because in such cases the assumption is that the supplementary test itself is free of bias and that it is testing the same construct as the test of focus (Bachman, 2004; McNamara & Roever, 2006).

DIF is not evidence for bias in the test. It is evidence of bias only if the factor causing DIF is irrelevant to the construct underlying the test. If that factor is part of the construct, it is called *impact* rather than bias. The decision as to whether the real source of DIF in an item is part of the construct being gauged is totally subjective. Usually, a panel of experts is consulted to give more validity to interpretations. As will be discussed later in the paper, however, the ad hoc nature of such interpretations have proved to be problematic.

3. Problems facing DIF research

In this section, a brief overview of the current problems in DIF analysis is presented. The discussion will specifically focus on four problems that the researchers consider to be of utmost importance: (a) the inter-method indeterminacy, (b) the intra-method indeterminacy, (d) the ad hoc interpretations, and (d) the impact of DIF on validity. The implications of each problem for current research and practice will also be given due attention.

3.1. Inter-method indeterminacy

The first problem facing DIF analysis dealt with here is the indeterminacy of the method used for DIF detection. As stated earlier, there is a plethora of techniques suggested, ranging from the traditional item-difficulty based approaches to the sophisticated statistical techniques such as Item Response Theory (IRT) and even Structural Equation Modeling (SEM). The problem with so many suggested methods is the often conflicting results that they produce (Bachman, 2004; McNamara & Roever, 2006).

Lei, Chen, and Yu (2006), for example, compared the performance of four DIF detection techniques: Mantel-Haenszel, SIBTEST, logistic regression, and IRT. The results indicated that the performance of these techniques were not comparable under different sample size ratios and impact conditions in terms of Type I error, power, and specificity in identifying the form of DIF. Rogers and Swaminathan (1993) compared the performance of logistic regression and Mantel-Haenszel. They reported that Mantel-Haenszel was not as powerful as logistic regression in detecting non-uniform DIF. The problem is more significant in practical settings. At times, applying different DIF techniques will identify different items as displaying DIF. For example, Karami and Shabani (forthcoming) have compared the performance of the Rasch model and Mantel-Haenszel in DIF detection. They report that only half of the items detected as showing DIF by Mantel-Haenszel were also flagged by the Rasch model. This is clearly a cause for concern. If the mere selection of the method exerts so much influence on the number of items flagged as DIF, then how sure can we be of the results of DIF analysis? Very often the decision as to what method to apply is quite haphazard. This is clearly an unfavorable situation in DIF analysis. The issue will be taken up again later.

3.2. Intra-method indeterminacy

In addition to the indeterminacy between the techniques for DIF detection, a similar situation exists when we come to each individual method. There are two notable problems here.

The first problem pertains to the existence of variations of the same technique. Take Mantel-Haenszel as an example. There are different types of this method suggested in the literature. Sometimes, the performance of these methods are not comparable. For example, Penfield (2001) compared three variations of the Mantel-Haenszel: the Mantel–Haenszel with no adjustment to the alpha level, the Mantel–Haenszel with a Bonferroni-adjusted alpha level, and the Generalized Mantel–Haenszel (GMH) that offered a single test of significance across all groups. Much variation was observed in the performance of these methods under different conditions including sample size, focal group ability distribution, and magnitude of matching criterion contamination. This adds to the inter-method indeterminacy discussed earlier.

The second problem is related to the kind of rules offered for interpreting the results of DIF analysis within each method. Take the Rasch model as an example. Rasch analysis software such as Winsteps (Linacre, 2010a) calculate DIF and offer a significance level. The significance level shows that the difference between the performance of the groups on the item is significant. Therefore, a significant difference at *p<05* level is a *prima facie* evidence for DIF. That is, the item is displaying DIF at *p<05* level.

This way of interpreting the results, however, is not taken up by all scholars. Some researchers such as Linacre (2010b) suggest that DIF contrasts (differences between item difficulty estimates for two groups) smaller than 0.5 are not practically significant. Others (e.g. Pallant & Tennant, 2007) have applied a Bonferroni adjusted alpha level without considering the DIF contrast at all. The ironic point is that the researcher will come up with different numbers of DIF items with each approach. So much variation in the kind of methods applied and the intra-method variation is not desirable if DIF analysis aims to realize its full potential.

3.3. Ad hoc interpretations

The next problem pertains to the kind of interpretations offered for the sources of DIF. This is a very important issue because the whole value of DIF depends on this phase of the analysis. It was stated earlier that DIF is not necessarily an indication of bias in the test. Bias is ensured if, and only if, the factor or factors bringing about DIF are not part of the construct. Usually, a panel of experts convene to decide on the sources of DIF, i.e., to determine whether DIF is a sign of bias or not. The whole process is completely subjective. This is a cause for concern because there is often no agreement among the experts as to the real source of DIF. If the experts do not reach an agreement, then it is not clear whose judgment should be taken at face value. Even if they reach an agreement, there is no guarantee that the judgment is true.

The kind of interpretations offered in DIF analysis are what Alavi and Karami (forthcoming) have dubbed as "ad hoc." The term ad hoc is used in a special sense, the way Popper intended it (Popper, 1934). In his special view of science, known as falsificationism, Popper was mainly concerned with all-encompassing theories that seemed to explain everything (Ladyman, 2002). That is, any evidence appeared to confirm the predictions of the theory and none to refute it. Even if some contrasting evidence was offered, there was a swift response on the part of the proponents of the theory adding some extra conditions to their theory in order to save it from rejection. These interpretations Popper called "ad hoc." The interpretations offered for the possible causes of DIF are ad hoc because they are not refutable by the existing evidence. No evidence exists to either confirm or disconfirm such interpretations. They are just hypotheses proposed and, in the absence of supporting evidence, may lack the real scientific vigor expected (Popper, 1934).

Bond (1993) recites an experience of working as a graduate student trying to provide explanations for the detected DIF items. It nicely depicts the kind of ad hoc interpretations intended here.

> She and I spent the better part of an afternoon devising elaborate and ostensibly convincing theories about why six particular items on the Metropolitan Achievement Test were behaving differentially for Black examinees, only to discover that, because of a programming error, we had been examining the wrong items. What was especially painful was the realization that in subsequent theorizing about the correct set of items showing Differential Item Functioning (DIF), we found ourselves making arguments that were diametrically opposed to our earlier theorizing, (p. 277).

In order to show the ad hoc nature of such interpretations, Alavi and Karami (forthcoming) asked a panel of experts to comment on the possible sources of DIF in a set of items. The informants were all selected from among TEFL graduates who had done scholarly research in language testing and were familiar with DIF analysis. There were two sets of items given to the informants: those showing DIF in favor of the Science and Technology students, and those which displayed DIF in favor of Humanities students.

There were six items in each set. In each set, two items really displayed DIF in favor of the identified group, two items had been found to display DIF against that group but for the purposes of the study, the informants were told that they were favoring this group, and finally, two items that had shown no significant DIF whatsoever. All these six items were included in a set in a random order and the informants were informed that they had all displayed DIF in favor of one group, either Humanities or Science and Technology (Alavi and Karami, forthcoming).

The results of that study were in fact intriguing. In fact, there was no order to the informants' interpretations and they resorted to different strategies to justify the existence of DIF. The majority of them focused on the relevance of the stems of the items to a specific grouping. However, when the stem was not relevant in that way, they changed their focus from the stem to the alternatives and their relevance to the examinees' background. The researchers concluded that, if they had learnt anything from their study, it was the fact that there was no agreement among the experts as to why DIF had occurred (Alavi and Karami, forthcoming).

This problem is not a mere theoretical debate. It is at the crux of DIF analysis. If DIF is a conclusive evidence for bias only if the underlying factor is not part of the construct and if we cannot determine, with some certainty, whether it is, then what is the use of DIF analysis at all. Determining the source of DIF is the final arbiter that determines whether the item is biased or not. All the value of DIF analysis depends on this final stage. DIF studies will be of little theoretical plausibility if some order is not brought to these interpretations.

3.4. Impact of DIF on validity

One of the central issues in DIF analysis is the examination of the impact of DIF on test validity. A number of research studies (e.g., Roznowski & Reith, 1999; Zumbo, 2003) have attempted to statistically model the impact of DIF on test performance. The results, however, have been mixed. While some researchers such as Roznowski and Reith (1999) and Zumbo (2003) have reported that DIF has little, if any, impact, Pae and Park (2006) report that DIF may affect the performance on the test. The issue is of much significance because, as Pae and Park (2006) state, "it can provide new insights into how DIF items in the item bank should be dealt with, and because decisions with a test are made not by the result of an individual item score but by the result of a whole test score" (p. 476).

Such discrepancies in research findings, though problematic in and of themselves, are not the central issue. The more important point is that regardless of their findings, these studies have focused on the impact of DIF on the mean performance of a group of examinees. Suppose that there is a language proficiency test with a cut-score of, say, 60. Suppose further that there are a large number of DIF items in the test favoring either males or females. As usual, a test fairness analysis is undertaken by comparing the performance of the two groups on the original test and an item composite comprised of only neutral items.

Assuming that the mean performance of the two groups did not differ in the two tests, the researcher would conclude that there is no bias in the test and validity is not undermined by any means. The problem arises when we consider the performance of the individuals rather than the groups. What if an individual has scored 59 on the test? He would certainly fail the test because the cut-score is 60. Would that person fail the test if there weren't so many DIF items in the test? Wouldn't that person be able to get at least one item correct and pass the test if so many items did not disfavor his group, or show DIF in favor of the opposite group? These are important questions that cannot be answered by a mere comparison of the overall performance of the two groups on the original test and another test made up of only neutral items.

4. Conclusion and future directions

In this section, some suggestions are presented for future research. The focus will be on the problems discussed in this paper.

It was argued earlier that the inter- and intra-method indeterminacies have brought about a situation where the selection of the techniques and the kind of rules offered for DIF detection in each method are exerting much influence on the number of items flagged as showing DIF. The bottom line here is that researchers should not put all their eggs in just one basket. Applying only one method for DIF detection and totally relying on the results of just one method may not be justified in face of the problems just discussed. Therefore, it is suggested that more than one technique be applied in any DIF analysis. Though not a panacea, it provides a mechanism for selecting items that have been identified as showing DIF by more than one method. If an item is flagged by more than one method, we have more justification for regarding it as displaying DIF.

The next problem discussed here was that of the ad hoc interpretations. One possible direction for future research has been recently pointed out by Ercikan, Arim, Law, Domene, and Lacroix (2010). These researchers have exploited think aloud protocols (TAPs) to confirm the interpretations of DIF made by a panel of content experts. TAPs confirmed the interpretations of the experts for only 10 out of the 20 items included in the test. Ercikan *et al.*

(2010) took this to indicate that "evidence from expert reviews cannot be considered sufficient in deciding whether DIF items are biased and judgments about bias in test items need to include evidence from examinee thinking processes" (p. 33). Though of much significance to DIF analysis, further research is needed before use of TAPs realize their full potential.

As for the impact of DIF on test validity, it may be suggested that validity is totally context-dependent. We cannot make resort to a mere comparison of ability estimates on two tests, as is the current practice among researchers, and then claim that validity is not undermined. In high-stakes tests, and especially when there are cut-scores, it is incumbent on the test users to investigate the impact of DIF items on the individual examinees' test scores rather than those of the groups. DIF items may not affect the validity of the test for different groupings but they may affect validity when individuals are considered. Every attempt should be made to ensure that no one is unduly affected by the existence of DIF items in the test.

The Authors

Hossein Karami (hossein.hkarami@gmail.com) has received an MA in TEFL from the Faculty of Foreign Languages and Literature, University of Tehran, Iran. His research interests include language testing in general, and Differential Item Functioning, validity, and fairness in particular.

Mohammad Ali Salmani Nodoushan (Salmani.nodoushan@yahoo.com) has received his PhD in Applied Linguistics. He has published several papers in international scholarly journals including *Teaching and Teacher Education*, *Speech Communication*, *TESL Canada Journal*, and so on.

References:

Alavi, S. M., & Karami, H. (Forthcoming). Differential Item Functioning and ad hoc interpretations.

Alderson, J. C., & Urquhart, A. (1985). The effect of students' academic discipline on their performance on ESP reading tests. *Language Testing*, *2,* 192-204.

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.

Bond, L. (1993). Comments on the O'Neill & McPeek paper. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 277–280). Hillsdale, NJ: Lawrence Erlbaum Associates.

Brown, J. D. (1999). The relative importance of persons, items, subtests and languages to TOEFL test variance. *Language Testing*, *16*, 217–238.

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

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.

Ercikan, K., Arim, R., Law, D., Domene, J., & 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*, 24–35*.*

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.

Hale, G. A. (1988). Student major field and text content: interactive effects on reading comprehension in the Test of English as a Foreign Language. *Language Testing, 5*, 49–61.

Kamata, A., & Vaughn, B. K. (2004). An introduction to differential item functioning analysis. *Learning Disabilities: A Contemporary Journal*, *2*, 49-69.

Karami, H. (2010). A Differential Item Functioning analysis of a language proficiency test: An investigation of background knowledge bias. Unpublished Master's Thesis. University of Tehran, Iran.

Karami, H. (2011). Detecting gender bias in a language proficiency test. *International Journal of Language Studies, 5*(2), 167-178.

Karami, H., & Shabani, E. A. (2011). On the comparability of two DIF detection techniques: Mantel-Haenszel and the Rasch model. Paper to be presented at the upcoming TELLSI conference (October 20-22), Ilam, Iran.

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

Ladyman, J. (2002). *Understanding philosophy of science.* London: Routledge.

Lei, P. W., Chen, S. Y., & Yu, L. (2006). Comparing methods of assessing Differential Item Functioning in computerized adaptive testing environment. *Journal of Educational Measurement, 43*(3), 245-264.

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., & Roever, C. (2006). *Language testing: The social dimension*. Malden, MA & Oxford: Blackwell.

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

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*(2), 13-23.

Pae, T. (2004). DIF for learners with different academic backgrounds. *Language Testing, 21*, 53–73.

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

Pallant, J. F., & Tennant, A. (2007). An introduction to the Rasch measurement model: An example using the Hospital Anxiety and Depression Scale (HADS). *British Journal of Clinical Psychology, 4*, 1–18.

Penfield, R. D. (2001). Assessing differential item functioning among multiple groups: A comparison of three Mantel-Haenszel procedures. *Applied Measurement in Education, 14*, 235–260.

Popper, K.R. (1934). *The Logic of Scientific Discovery*. London: Hutchinson.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement*, *17*(2), 105-116.

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.

Ryan, K., & Bachman, L. (1992). Differential Item Functioning on two tests of EFL proficiency. *Language Testing, 9,* 12–29.

Shabani, E. A. (2008). Differential Item Functioning analysis for dichotomously scored items of UTEPT using Logistics Regression. Unpublished master's thesis, University of Tehran, Iran.

Takala, S., & Kaftandjieva, F. (2000). Test fairness: A DIF analysis of an L2 vocabulary test. *Language Testing, 17,* 323–340.

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, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.

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