Interpreting the relationship between item difficulty and DIF: examples from educational testing

M Wilson and M Duhlunty

1 Graduate School of Education, University of California, Berkeley, USA
2 Melbourne Graduate School of Education, University of Melbourne, Australia

E-mail: MarkW@berkeley.edu

Abstract. The investigation of differential item functioning (DIF) is commonplace in educational testing. Most often it results in no more than the discarding of a proportion of the developed items. Apart from the potential threat to validity of the instrument due to differential loss of certain types of items, this also is inefficient in terms of item development costs. This paper discusses certain situations where particular patterns of non-uniform DIF can be related to item difficulty, and modelled by enhanced Rasch models. This allows for a more efficient usage of items, as well as, potentially, an improved understanding of the complexities of how educational tests function across different sub-populations.

1. Introduction

In educational testing, as in other measurement contexts, the validity of interpretation of the results of statistical analysis of measurement data depends critically on the reasonableness of the assumptions behind the statistical models used to analyze the data. If these assumptions are violated, then serious mistakes may result from the uncritical usage of the measurements. One particular family of such threats to validity is differential item functioning (DIF), which occurs whenever the probability of a particular response to a specific item is influenced by factors other than the respondent’s underlying ability [1-3].

While this may occur at the individual student level (as it almost certainly does), the existence of such effects is most readily detected at the group level—standard cases would be differentials between ethnic and racial groups, socio-economic groups, or gender effects. For example, in US contexts, comparing African-American students with White students is quite common, and where certain items are found that display evidence of DIF, the test developers systematically delete them from the test. This is done usually without any recourse to investigating why there was DIF for that particular item for that particular group comparison, as the existence of DIF for an item is seen as indicating that the item is in some way unfair, and hence must be eliminated. However, this systematic deletion of items from the item sample may constitute a threat to test validity in itself (in the sense of effectively altering the definition of the underlying property under measurement), especially where the deletions were concentrated on specific types of items etc.

Thus, it is of interest to try and find patterns in DIF results, and then relate these patterns to substantive aspects of the underlying property and/or features of the measurement instrument. The aim of this paper is to examine one particular way that DIF may be systematically occurring in tests, to relate that to potential patterns of DIF results, to describe some statistical models that could be used in this situation, and to recommend next steps in a research agenda focused on this issue.
2. Relating DIF to item difficulty

One prominent example of systematic DIF was brought to attention by Freedle [4] who showed how many of the more difficult SAT items displayed DIF effects which favored African American students while easier SAT items had DIF effects favoring White students. Freedle hypothesized that this relationship between item difficulty and DIF estimates could be explained by cultural familiarity and semantic ambiguity. He postulated that, on the one hand, "Easy" Verbal items are contextualized within a more culturally specific content and hence can be hypothesized to be perceived differently, depending on the student’s own cultural and socioeconomic background. On the other hand, "Hard" Verbal items will tend to be expressed using rarely used words that are less likely to have differences in interpretation across ethnic communities because these are only familiar to those with higher levels of education, which is more uniform than the home environment.

This explanation leads then to an idea that, in this case, for the comparison of African-American students and White students, there might be a monotonic relationship, or possibly a linear one, between DIF and item difficulty, with the DIF effect decreasing as the difficulty increases. Indeed, this is similar to what Freedle in [4] found in his analyses, and he reported a correlation approximately of 0.5. Similar results in similar situations have been reported by Kulick and Hu [5], Scherbaum & Goldstein [6] and Santelices and Wilson [7]. Interestingly, these have all been found to be approximately linear relationships.

There has been considerable criticism of the interpretation of these results. Some researchers have claimed that the results are due to one of the following explanations: (i) the confounding of DIF and impact by the shortcomings of one particular technique called the Standardization Approach and/or (ii) by random guessing. These researchers were all associated with testing companies however [8-12]. An investigation by researchers not associated with a testing company found that the results persisted even when these effects were accounted for [13].

3. Statistical models related to DIF

In a standard dichotomous Rasch model,

\[ \text{logit}(X_{pi} = 1 \mid \theta_p, \delta_i) = \theta_p - \delta_i, \]  

where \( X_{pi} \) is the response of person \( p \) to item \( i \) (which is assumed to be dichotomous, so that \( X_{pi} \) has two values 0 and 1), \( \theta_p \) is the person location, and \( \delta_i \) is the item difficulty for item \( i \).

Now, assume there is a student covariate, \( C_p \), which could be, for example, a student’s gender or ethnicity. For simplicity’s sake, assume that \( C \) is dichotomous—i.e., it takes the values of either 1 or 0 — hence, we can form the labeling function \( c(p) \), which takes on the value 1 when \( p \) is in group \( C_p \), and 0 otherwise. In the psychometric literature the group with \( c(p) = 1 \) are termed the focal group, and the group with \( c(p) = 0 \) are termed the reference group. The standard Rasch DIF model [14] is given by

\[ \text{logit}(X_{pi} = 1 \mid \theta_p, \delta_i, \gamma_i) = \theta_p - \delta_i + c(p)\gamma_i, \]  

where \( \gamma_i \) is the DIF effect for item \( i \). This model is not estimable unless we impose a constraint, for example, that \( \gamma_0 = 0 \). This constraint is often expressed in terms of the reference group—the difficulty parameters are seen to be those for the reference group, and the DIF parameter expresses how much the item difficulty for the focal group differs from this. This is referred to as having uniform DIF.

A linear relationship between item difficulty and the DIF effect, as postulated above, could then be expressed as:

\[ \text{logit}(X_{pi} = 1 \mid \theta_p, \delta_i, \gamma_i) = \theta_p - \delta_i + \theta_p c(p)\gamma_i. \]  

Or, alternatively, using a function that is monotonically-related to person ability, such as the sufficient statistic for the person ability under the Rasch model, the person score\(^1\), denoted as \( t(p) \).

---

\(^1\) Under an assumption of no missing data.
of the "DIF plots" in Dulhunty [16] were similar to those in figure 3. (up to approximately 0.5 logit)  

Example of an absence of a linear DIF effect versus non is sown in figure 3, which plots DIF effects against item difficulties in the Spelling test for Indigenous using equation 2 above, but is a possible confounding effect when two separate calibrations are each of the referen
differential shrinkage of the item difficulties in the separate calibrations carried out by NAPLAN for investigation of an alternative explanation in the case of Gend
g years, and the items in the NAPLAN tes
ing indigenous students, as the phenomenon of gender bias in educational tests has been known for many it would seem less likely that this could be due to a similar "cultural familiarity" effect as for the Australian Male students at the lower end of the ability spectrum find the items more difficult than do their own complexities. 

4. Examples 

For example, recent work by Dulhunty [16], illustrated in figure 1, shows evidence for a systematic relationship between DIF related to indigenous status among Australian students and item difficulty in the national reading test for Grades 3 and 5 in 2017. These results are expressed in terms of item-wise DIF estimated using equation 2 with non-Indigenous students used as the reference group. The reported results were obtained by comparing an item calibration using only the reference group (i.e. non-Indigenous students) to a separate item calibration using only the focal group (i.e. Indigenous students). Each point on the plot refers to a specific test question. Negative y-axis values indicate that the item is relatively easier for Indigenous students (after adjusting for average ability). Positive values indicate Indigenous students find the item relatively difficult compared to non-Indigenous students. These figures display a similar pattern as was observed by Freedle for the comparison of African-American with White students in the US. The interpretation by Freedle could be applicable here also—Australian Indigenous students at the lower end of the ability spectrum find the items more difficult than do their equal-ability non-Indigenous students, while the effect reverses for higher ability Indigenous students.

When we look at the DIF effects for gender in the same test, similar results are also found for Grades 3 and 5 (see figure 2). In this case, consistent with the comments above, the interpretation would be that Australian Male students at the lower end of the ability spectrum find the items more difficult than do their equal-ability Female students, while the effect reverses for higher ability Male students. However, it would seem less likely that this could be due to a similar “cultural familiarity” effect as for the indigenous students, as the phenomenon of gender bias in educational tests has been known for many years, and the items in the NAPLAN test have been screened for this. Thus, there needs to be investigation of an alternative explanation in the case of Gender—a possibility would be that there was differential shrinkage of the item difficulties in the separate calibrations carried out by NAPLAN for each of the reference and focal groups. Note that this effect would not be present in analyses carried out using equation 2 above, but is a possible confounding effect when two separate calibrations are used (and this is one reason why equation 2 was used to define the DIF effect in the account above).

Now, it is important to examine instances of linear DIF effects as well as non-instances. One example is sown in figure 3, which plots DIF effects against item difficulties in the Spelling test for Indigenous versus non-Indigenous students in the same grades and year as in figures 1 and 2. Here we see an example of an absence of a linear DIF effect—there are indeed largish DIF effects for individual items (up to approximately 0.5 logit)—but the plots are all centered around a DIF effect of 0.0. In fact, most of the “DIF plots” in Dulhunty [16] were similar to those in figure 3.
Figure 1. Plot of DIF effects with item difficulties when comparing Australian Indigenous and non-Indigenous students for Grades 3 and 5 Reading.

Figure 2. Plot of DIF effects with item difficulties when comparing Australian Male and Female students for Grades 3 and 5 Reading.

Figure 3. Plot of DIF effects with item difficulties when comparing Australian Indigenous and Non-indigenous students for Grades 3 and 5 Spelling.
5. Discussion
The analysis of publicly reported results for 240 NAPLAN test forms [16] shows DIF estimates can at times be found to be correlated with item difficulty. Furthermore, the strength of the relationship depends on the assessment domain for which DIF was examined and the population subgroups for which the DIF analyses were performed. Some relationships appear to be stable year after year despite the use of different test forms administered to different student cohorts. For example, there appears to be a linear relationship between Rasch estimated item difficulty and Indigenous DIF on the Reading assessments. The association persists across year levels (from Year 3 to Year 9) and has been observed annually from 2013 to 2017 (which are all the technical reports publicly available to date).

The results reported in [16] are based on the assumption inherent in the NAPLAN analyses that the metrics resulting from the two Rasch model calibrations (i.e., for the reference group and the focal group) are the same. Preliminary investigations suggest this assumption is not always valid: When calibrated separately the standard deviations of the item parameters are sometimes not the same. When this occurs we end up with the familiar pattern of one group finding easy items relatively hard and, at the same time, finding hard items relatively easy. This phenomenon can be considered a case of test equating when a ‘shift and scale’ adjustment is required to equate tests rather than just a ‘shift’ (e.g., [17]). This type of issue has been recognized as an issue for measurement in general [18]. Further investigation of the NAPLAN results (say, using equations 2 to 5) are needed to eliminate the possibility that some observed linear DIF effects may be the result of a scaling artifact.

6. Conclusion
The phenomenon of item-wise DIF is a common occurrence in educational testing: it mainly constitutes a headache for test developers, who typically develop extra items beyond those that are expected to be needed specifically to account for loss of items due to DIF. As mentioned above, although this practice does indeed result in a set of items that are not subject to DIF effects for the groupings that have been tested, the elimination of particular types of items could be a source of bias in the item pool. Fitting models such as those in equations 3 to 5 can allow the possibility that items with these specific DIF patterns can be maintained within the item pool, and thus avoid such threats to validity.

The accumulation of results such as these leads one to seek more robust theoretical understanding of the phenomenon. The models in equations 4 and 5 allow an efficient and unbiased estimation of the underlying relationship (when the assumptions are correct), and also the estimation of bias-corrected measurements of students in both groups. Each time this phenomenon is found, it potentially contributes to a deeper understanding of the complications of testing across different sub-populations.

Further work along these lines is needed. Of course, one important step is to go beyond the examination of reported DIF effects as described above, to the direct estimation of linear DIF effects utilizing equations 3 to 5. This will require a careful re-analysis of the data at the item-level, and, as mentioned above, the investigation of appropriate estimation software for that purpose (and investigation for any scaling artifacts using, say, simulation studies). Beyond that, interpretation of what it would mean if linear DIF effects were indeed confirmed, then further investigations into the possible causes for such effects would also be needed—these could include (a) deep dives into the relevant literature (which would vary, of course, depending on the reference/focal group pair), and (b) simulation studies to eliminate possible statistical artifacts, as well as (c) qualitative investigation of student response patterns.

References
[1] Lord F M 1980 Applications of Item Response Theory to Practical Testing Problems (Hillsdale, New Jersey: Lawrence Erlbaum Associates)
[2] Hambleton R, Swaminathan H and Rogers J 1991 Fundamentals of Item Response Theory (Thousand Oaks, California: Sage Publications)
[3] Thissen D, Steinberg L and Wainer H 1988 Use of item response theory in the study of group differences in trace lines Test Validity ed H Wainer and H Braun (Hillsdale, New Jersey: Lawrence Erlbaum Associates)

[4] Freedle R 2003 Correcting the SAT’s ethnic and social-class bias: A method for re-estimating SAT Scores Harvard Educ. Rev. 73 1–43

[5] Kulick E and Hu P G 1989 Examining the Relationship Between Differential Item Functioning and Item Difficulty (College Board Report No. 89-5; ETS RR-89-18) (New York: College Entrance Examination Board)

[6] Scherbaum C and Goldstein H 2008 Examining the relationship between race-based differential item functioning and item difficulty Educ. Psychol. Meas. 68 537–553

[7] Santelices M V and Wilson M 2010 Unfair treatment? The case of Freedle, the SAT and the standardization approach to differential item functioning Harvard Educ. Rev. 80 106–133

[8] Dorans N 2004 Freedle's Table 2: Fact or fiction Harvard Educ. Rev. 74 62–73

[9] Dorans N and Zeller K 2004 Examining Freedle’s Claims and His Proposed Solution: Dated Data, Inappropriate Measurement, and Incorrect and Unfair Scoring (Educational Testing Service Report No. RR-04-26) (Princeton, New Jersey: Educational Testing Service)

[10] Bridgeman B and Burton N 2005 Does scoring only the hard questions on the SAT make it fairer? The Annual Meeting of the American Educational Research Association (Montreal)

[11] Wainer H 2009 Ethnic bias or statistical artifact? Freedle’s folly Picturing the Uncertain World. How to Understand, Communicate and Control Uncertainty through Graphical Display ed H Wainer (New Jersey: Princeton University Press)

[12] Wainer H and Skorupski W 2005 Was it ethnic and social-class bias or statistical artifact? Logical and empirical evidence against Freedle’s method for reestimating SAT scores Chance 18 17–24

[13] Santelices M V and Wilson M 2012 On the relationship between differential item functioning and item difficulty: An issue of methods? Item response theory approach to differential item functioning. Educ. Psychol. Meas. 72 5–36

[14] Meulders M and Xie Y 2004 Person-by-item predictors. Explanatory Item Response Models. A Generalized Linear and Nonlinear Approach ed P De Boeck and M Wilson (New York: Springer)

[15] Wu M, Adams R J and Wilson M 1998 ACER-ConQuest (Hawthorn, Victoria, Australia: ACER Press)

[16] Dulhunty M 2019 Item Difficulty Association with Reported Gender, Indigenous and LBOTE DIF in NAPLAN 2013 to 2017 (Melbourne, Australia: MGSE, University of Melbourne)

[17] Humphry S M 2011 The role of the unit in physics and psychometrics Measurement 9 1–24

[18] Wu M, Hak P T and Tsung-Hau J 2016 Educational Measurement for Applied Researchers: Theory into Practice (Singapore: Springer)