Journal of Education and Human Development
June 2014, Vol. 3, No. 2, pp. 753-774
ISSN: 2334-296X (Print), 2334-2978 (Online)

Published by American Research Institute for Policy Development

Differential Item Functioning and its Utility in an Increasingly Diverse Classroom: Perspectives from Australia

Alvin Vista, PhD[1] & Esther Care, PhD[2]

Background

This paper discusses the utility of conducting differential item analysis on standardized tests and the implication of such analysis in today's increasingly diverse classrooms. This research into practice discussion is set in the context of government schools in Victoria whose student population is becoming multi-cultural and multi-ethnic. Victoria, and Australia as a whole, is experiencing rapid changes in its population demographics. Almost half (49%) of first-generation Australians now speak a language other than English at home, and up to 20% even for second-generation Australians (Australian Bureau of Statistics, 2006). In Victoria, approximately 20.4% of the population is now from non-English speaking backgrounds (NESB) (Department of Immigration and Citizenship, 2008). One in four (24.7%) people within the Melbourne statistical division is classified as a speaker of language other than English (LOTE) (Department of Justice, 2005). It is in this context of diversity that this paper is set.

The main aims are to present the problems involved in the use of uniform assessment within diverse classroom populations and to discuss the utility of differential item functioning (DIF) analysis as a critical tool in test development. This paper is targeted for practitioners and thus presents DIF analysis techniques that are practical for actual school-level (or even classroom-level) implementation and within the statistical knowledge of most school personnel.Five DIF analysis methods of varying complexity and level of utility are presented, with the goal that the readers will be able to choose the most appropriate method to analyse DIF for their respective contexts.

[1] Research Fellow | Assessment Research Centre| Melbourne Graduate School of Education | The University of Melbourne | VIC 3010 Australia.
[2] Associate Professor | Melbourne Graduate School of Education | The University of Melbourne | VIC 3010 Australia

In proposing that practitioners conduct DIF analysis on their own data, the authors anticipate that by becoming familiar with the overall issue of DIF and having some capacity to identify it, practitioners also become aware of the deeper issues concerning test bias and equity.

Because of this relationship between DIF and test bias, including the implications for equity, this paper is also useful for policy makers even if they may not be directly involved in conducting DIF analyses.

DIF analysis is an important part of a larger system of inquiry into issues of equity in assessments. In turn, issues of equity in assessment are linked with the broader issue of educational equality (e.g., Gipps & Murphy, 1994). While the issues of equity in assessment involve the whole educational evaluation process, DIF analysis is mainly focused on formal large-scale assessments and works best with quantitative, objective, and standardized tests (McNamara & Roever, 2006).Typically, these types of tests are regarded as'high stakes'tests which may have impact on test-takers' educational future. Accordingly, among the various tools that can be used to detect test bias, DIF analysis is among the most critical for detecting bias that may affect educational futures.

Australia is moving towards a national testing standard to assess performance in schools. Whereas in the past, the states have always developed their own curriculum and forms of assessment, Australia is now developing a single national curriculum and a uniform set of standardized tests to assess the students' performance. The National Assessment Program – Literacy and Numeracy (NAPLAN)has been administered nationally since 2008 to assess literacy (specifically Reading, Writing, and Language Conventions) and numeracy skills at year levels 3, 5, 7, and 9 as part of the thrust towards a standardAustralian curriculum (MCEECDYA, 2008)

What is DIF?

DIF can be defined as a statistical phenomenon that occurs when "two individuals with equal ability but from different groups do not have equal probability of success on the item" (Shepard, Camilli, & Averill, 1981, p. 319).In other words, DIF occurs when examinees from different groups show differing probabilities of success on (or endorsing of) the item after matching on the underlying ability that the item is intended to measure.

The relationship between the differences in probabilities and the underlying ability of one of the groups dictates the type of DIF – uniform or non-uniform.

DIF is defined as *uniform* is the differences in probabilities are uniform (e.g., one group retains the advantage) over all ability levels; it is *non-uniform* if the difference in the probability of success changes depending on ability level (e.g., the advantage switch from one group to the other for low vs high ability level) (Swaminathan & Rogers, 1990). To detect the existence of DIF, and to a lesser degree determine the type, we need to conduct DIF analysis, of which there are several methods that will be presented here.

DIF analysis therefore is a statistical tool that can be used to improve how a test behaves across groups and to reduce group-based differentials that are not relevant to the construct being measured. In terms of utility, DIF analysis is mainly a tool to assess test fairness, investigate threats to validity, and explore the underlying processes in item responses across groups (Zumbo & Gelin, 2005). The relationship between DIF and test bias, implications for fairness, and consequences are discussed in the succeeding sections.

DIF and Test Bias

DIF is required, but not sufficient, for item bias (McNamara & Roever, 2006). Thus, bias is a broader term. If there is bias, there has to be DIF; but if there is DIF, it does *not* necessarily mean that there is bias (Clauser & Mazor, 1998). This distinction between DIF and item bias is defined more formally by Camilli(1993):

If the degree of DIF is determined to be practically significant for an item and the DIF can be attributed plausibly to a feature of the item that is irrelevant to the test construct, then the presence of this item on the tesbiases the ability estimates of some individuals. This compound condition, when satisfied, indicates item bias. (p. 398)

Item bias occurs when examinees of one group are less likely to answer an item correctly than examinees of another group because of some characteristic of the test item or testing situation that is not relevant to the test purpose. This requirement of relevance to the test purpose is essential because group differences in areas being measured can be interpreted as real differences rather than bias. Consider if instead of a maths test, we are examining the differences in an IELTS test.

It is likely that we will find evidence of DIF between students fromnon-English speaking backgrounds (NESB) and English speaking backgrounds (ESB) there, but that would not be considered as bias because English proficiency is what IELTS is supposed to measure. In fact, if there are no differences between NESB and ESB in IELTS, the validity of the test will be in doubt.

Validity is affected by construct-irrelevant variance, which is defined by Messick(1994) as occurring when a test measures something that is irrelevant to the construct of interest. For example, a timed computer-based essay-writing test would measure in part a test taker's typing proficiency and speed, which would be a form of construct-invariance since the main construct of interest is essay-writingskills. The issue of construct-irrelevant variance similarly becomes apparent when two groups of examinees are tested on their proficiency in aconstruct that does not involve reading comprehension (for example, mathematics) but whose performance on that construct is nevertheless significantly influenced by the groups' reading comprehension skills. Where construct-irrelevant variance differs across groups, we find another case of DIF. This concern is examined in a large-scale study involving student data from the Trends in International Mathematics and Science Study (TIMSS), which used logistic regression analysis to detect uniform and non-uniform DIF between groups based on primary language spoken at home (Hauger & Sireci, 2008). The findings were positive for TIMSS as no DIF with practical significance was found among the test items, indicating that language proficiency is not interacting with the construct being measured (Hauger & Sireci, 2008). Further, evidence of this kind can be taken to assess the quality of test and item development, which underscores the importance of conducting DIF analysis on high-stakes tests.

Consequences of DIF

In the Unites States (US), where high-stakes tests are often used as a significant part of the admissions process to higher education the public has demanded legislated policies that force testing institutions to make their tests more rigorous in terms of technical and psychometric properties (Gipps & Murphy, 1994). The issue of equity in this context first revolved around group differences and allegations of test bias between Caucasians and African-Americans.

Due to several landmark court decisions and support from the professional measurement community, test companies began to collect data and use DIF analysis methods specifically to identify test bias (Gipps & Murphy, 1994). Items that function differently for disadvantaged groups can have far-reaching repercussions that increase in significance in proportion to the magnitude of stakes involved. This is one of the main reasons why legislated policies have become more prominent in the US.

The situation in Australia is slightly different to that in the US for a few important reasons: there is no comparable high-stakes test or set of tests that are used for university admissions; most large-scale state or nationwide tests are developed and administered by the government rather than private companies; and private schools (specifically the Catholic and independent school sectors) have a considerably larger share of the school population (and consequently hold greater influence) than in the United States (Broughman, Swaim, & Keaton, 2009; Ryan & Watson, 2004).In Australia, the task ofensuring that high-stakes tests remain free of significant DIF and bias rests with the respective federal and state departments, who are the main developers and implementing agency of these types of tests. Nevertheless, at the local, classroom level, the consequences of DIF on equity-equality issues remain relatively the same in Australian and US schools. In the local setting (at classroom or school level) and involving school practitioners, DIF analysis will be implemented mostly with relatively few test takers and with tests that have a lower range of statistical reliability compared to large-scale standardised tests. In addition, since the sensitivity of different DIF methods to group differences varies considerably, schools with large variability in student demographics also have greater discretion in choice of suitable DIF methods as well. This can give individual schools the flexibility to adapt different DIF detection methods to suit their specific needs.

Methods for DIF Detection

There are a number of DIF analysis methods but this paper will only present and discuss those that are accessible to most practitioners. Clauser and Mazor(1998) provide a more comprehensive overview of the statistical procedures to detect DIF; however, some DIF detection techniques are computationally intensive and conceptually complicated such that their utility is in the main limited to psychometricians.

Different DIF detection methods exist for a variety of test types. The test type can roughly be classified as those with dichotomous (only two possible scores) or polytomous(interval scores are possible) item formats, while the DIF approach can be classified as parametric or nonparametric (in our context, the main difference between them is that nonparametric approaches require no assumptions that the data follow a certain probability distribution). In this paper, the focus is on DIF methods, parametric and nonparametric, that work best for tests with dichotomous items (i.e., items that are scored as simply correct or incorrect). Thus, some approaches, such as the nonparametric standardized mean difference (SMD) index and cumulative common log-odds ratio, which are designed for polytomous item format (Penfield, Giacobbi, & Myers, 2007) are not included here.

To illustrate these methods, a dataset using test results from 187 preparatory level (Prep) children in Melbourne was used. This dataset is derived from a study that looked at the relationship between linguistic background and performance on common tests of ability used in Melbourne schools (Care, Roberts, & Thomas, 2009). Test data from 187 Prep children on the "Space concept" subtest of the Boehm Test of Basic Concepts, 3rd Edition (BOEHM-3, Boehm, 2001) will be used throughout this paper (see Care, et al., 2009 for details on data collection and methodology); with grouping based on linguistic background (A= ESB, B= NESB). In essence, the illustrations analyse the existence of DIF on a subtest of the BOEHM-3 between English and non-English speaking background Prep students.

The BOEHM-3 consists of 50 items that measure basic concepts such as quantity, space, and time (BOEHM-3, Boehm, 2001). In this paper, only the subset of items that measure the space concept are used and to maintain the consecutive numbering, the original items have been renumbered. The correspondence between the numbering used in this paper and the original numbering is given in Table 1.

Parametric Approaches

Analyses based on Item Difficulty

This approach compares item difficulty estimates based on the proportion of correct responses, or $p = \frac{\#\text{of correct responses}}{\#\text{of incorrect responses}}$.

The transformed item difficulty (TID) index, also known as the delta plot (Angoff, 1972) is a method that has been in use for several decades; it is one of the easiest to implement as well as being among the easier methods to understand conceptually. This method requires only the computation of item-difficulty or *p*-values for each of the groups to be compared. The computed *p*-values are then normalized (transformed into deltas) and plotted into a chart with the two axes representing each group being compared. This can be done easily in Microsoft Excel, as in Table 1, with results from the two groups A and B on the 25-item BOEHM-3 subtest. The computed deltas are standardized *p*-values for each item. The deltas are then plotted and a linear trend line is added, as shown in Figure 1. The distance from the trend line indicates the degree of differential functioning for each item, with the position on the left or right of the line indicating which group has the advantage for such item – in this illustration, items on the right side of the line are easier for group A. This plot showsvisually the potentially problematic items that function differently between two groups as indicated by the arrows pointing at items 17 and 25, while arrows pointing at items 6 and 10, which are included for comparative purposes, show that both items are nearer to the trend line. More formal methods to compute the degree of departure are available as described in Angoff and Ford (1974), but for this illustration, we used a computed distance from the trend line and a threshold of -1.0 to 1.0 as indicator of DIF (dashed lines in Figure 1). The delta plot method has undergone significant modifications through the years, with increasing complexity of testing situations demanding a corresponding increase in complexity and rigour of the method (see Angoff, 1982 for an overview on the more complex modifications of the delta plot method).

The delta plot method is comparatively simpler and less computationally intensive than the rest of the DIF detection methods presented in this paper. However, there is a valid concern that this method can confound item difficulty with item discrimination (Angoff, 1982), especially if the compared groups have significantly different performance levels in the construct being measured. For example, if group A performs significantly better than group B, an item that discriminates well on the test construct will be flagged by this method as having evidence of DIF. This concern can be minimized by matching the groups in terms of ability beforehand (Angoff, 1982), although there might be situations where that is not possible.

If the groups cannot be matched in terms of ability and there is reason to believe that significant differences exist, other methods for detecting DIF may be considered.

Item-Response-Theory-Based Approaches

These approaches include 1, 2, and 3-parameter IRT analyses. We will focus on the item characteristic curve (ICC) method using single-parameter (Rasch) model because this is the least computationally demanding IRT approach (Wu, Adams, Wilson, & Haldane, 2007) given the target application and audience of this paper. As an overview, this method involves items that are mapped on a uniform latent variable scale that fits into a single parameter IRT (Rasch) model. That is, the probabilities of success for the items, $p(X = 1)$, follow the following model:

$$p\left(X=1\right)=\frac{e^{(\theta_n-\delta_i)}}{1+e^{(\theta_n-\delta_i)}} \quad (1)$$

and the relationship between person ability and item difficulty is given by:

$$\theta-\delta=\log\left(\frac{p}{1-p}\right) \quad (2)$$

where θ_n represents the person ability, and δ_i represent item difficulty, both on the same scale (Wilson, Allen, & Li, 2006). Figure 2a is a visual representation of this model, with the X and Y axes representing ability and item difficulty respectively, and the Z axis representing $p(X = X = 1)$. Figure 2a shows that as ability increases, the chances of a correct response also increase while conversely, as item difficulty increases, the chances of a correct responses decrease. Because $\left(\frac{p}{1-p}\right)$ is the odds of a correct response, it can be shown that the difference in difficulty between two items remains the same regardless of the ability of the test takers, thus providing a way to estimate the item difficulty in a scale that is uniform and has an arbitrary point of origin across test takers (Kelderman, 1988).

In the same way, it can be shown that the estimated difference in ability between two test takers remains the same regardless of the item (see Irtel, 1995). Figure 2b illustrates this by showing that for any given X, the intervals in Y are uniform and vice versa. This is a property of the Rasch model called specific objectivity (Rasch, 1966, 1977). Due to this arbitrary nature of the scale, it can be set such that item difficulties are comparable across samples and test subgroups (Kelderman, 1988).

Model-fitting usually requires specific software and for this paper, ConQuest(Wu, et al., 2007) was used to demonstrate this method, although the software is capable of fitting other IRT models and not just the single parameter model. ConQuest can be used to test for DIF using the ICC method (see Wu, et al., 2007 for the operational details of the software) by fitting data to a Rasch model and plotting them with respect to a grouping variable into separate item characteristic curves. Figure 3 shows a hypothetical plot of the ICCs from two groups, with the vertical axis representing the probability of a correct response and the horizontal axis representing the ability level. This overlaid plot visually shows that the group represented by the red line has a greater chance of a correct response for any given ability level, and thus can be interpreted as evidence that this particular item exhibits DIF (Ironson, 1982). ConQuest allows for a chi-square test of parameter equality as well as fit statistics for each item (Wu, et al., 2007). It should be kept in mind however that chi-square as an absolute fit index tends to increase as sample sizes increase and thus are more likely to be statistically significant even if the differences are inconsequential (Bollen & Long, 1993; Scholderer, Brunso, Bredahl, & Grunert, 2004). This could have important consequences for interpreting the test results of parameter equality and fit statistics when the IRT approach is applied on tests with very large samples.

Conducting DIF analysis on our sample using ConQuest, the chi-square test of parameter equality is significant, $\chi^2_{(24)} = 37.51$, $p = .04$. A significant chi-square result is not in itself evidence of DIF (Wu, et al., 2007), but it should be taken as a sign to look deeper into what is causing the groups to perform differently. Looking into item-level results, some items appear problematic, showing differences between ESB and NESB(Table 4).

For example, group A performs 0.54 and .69logits higher than group B on items17 and 25 respectively, showing that these itemsaremore difficult for group B. in comparison, the difference between the two groups for items 6 and 10 is only 0.16 logits at maximum (Table 4). ICCs for groups A and B on items 17 and 25 are shown in Fig. 4.

Logistic Regression

Logistic regression involves calculating regression equation that predicts the probability of success on an item based on total score, group membership, and a product term that represents interaction between the total score and group membership. Because of the inclusion of the product term, representing interaction, it is capable of detecting non-uniform DIF, which can be considered as an advantage for this method (Swaminathan & Rogers, 1990). In a single predictor and single grouping variable situation, the general equation is:

$$\log\left(\frac{p}{1-p}\right) = b_0 + b_1 total + b_2 group + b_3\left(total * group\right) + \varepsilon \qquad (3)$$

The left hand term of this equation is the log-odds of getting a correct response, which is actually related to the IRT equation (Eq. 2) showing the relationship between person ability and item difficultyin the previous section.

If the group predictor is not significant, or if its effect size is small, it can be interpreted as insufficient evidence of DIF. One advantage of this approach compared to the IRT ICC method is that it can be implemented without specialized software. Logistic regression analysis can be implemented in Excel, although SPSS provides a less cumbersome platform.

In illustrating this approach, we used SPSS to conduct logistic regression on the same items of our dataset used in the preceding methods (6, 10, 17, and 25). To address the issue of multicollinearity regarding the product term (*total*group*), the data for predictor variables need to be centred (Howell, 2002).

This is done by subtracting the mean from each of the observations (i.e., $total_{centred} = total_i - total_{mean}$), which has the effect of reducing the correlation between the predictor variables and the product term while the original correlation between $total_i$ and $group_i$ remains the same (Howell, 2002).Thus, the predictors for the dichotomous outcome variable (items) become *total* $_{centred}$, *language*, and $total_{centred}$**language*. The results indicate that theregression coefficientsfor *language*is non-significant (all exp(*B*) = .67-1.79, *p*s> .10), providing no evidence of DIF for these particular items. If the regression coefficient of the grouping variable is significant, it will be necessary to look the regression coefficient of the product term – a significant product term coefficient would be an indicator of non-uniform DIF (Zumbo, 1999).

Nonparametric Approaches

Chi-square Procedure

Nonparametric approaches include chi-square and the Mantel-Haenszel(MH) odds ratio techniques. Both methods are less affected by differences in ability levels between groups, which can be considered an advantage compared to the delta plot method (Ironson, 1982). Only the simplest way to perform the full chi-square procedure will be presented here. This involves dividing the whole range of ability levels (based on total scores) into groups of ability. The number of ability groups is arbitrary although it is suggested that around 5 intervals be used (Scheuneman, 1979), more if sample sizes are large (Ironson, 1982). For each item, a 2 x 2 table of correct/incorrect values is constructed for every ability group. The chi-square statistic for each ability group is computed as:

$$\chi_i^2 = \frac{N_i(ad - bc)^2}{(a+c)(b+d)(a+b)(c+d)} \quad (4)$$

where*i* = ability group, $N_i = a + b + c + d$, and *a*,*b*,*c*,*d* are cell values of the 2 x 2 table representing the number of responses (Table 2).

The full chi-square statistic for an item is then computed as the sum of all the chi-square statistics for every ability group:

$$\chi^2_{\text{full}} = \chi_1^2 + \chi_2^2 + \ldots + \chi_i^2 \quad (5)$$

A full chi-square analysis of the sample data is beyond the scope and purpose of this paper. Instead, two items (17 and 25) that were flagged from the delta plot method will be used to illustrate both the chi-square and Mantel-Haenszelprocedures. For this illustration, 5 ability groups are used, based on total score percentile ranks (i.e., 20th, 40th, 60th, and 80th percentiles as cut off scores for the 5 groups). Table 2 presents the results from item 17, where $\chi^2_{\text{full}} = 0.17 + 2.11 + 0.63 + 0.62 + 0.18 = 3.72$, with the degree of freedom for significance testing being the number of ability groups (5 in this case). This value is not statistically significant, p= .59, and thus we cannot reject the null hypothesis for this particular item as exhibiting evidence of DIF. Table 3 presents the results for item 25, where $\chi^2_{\text{full}} = 0.85 + 6.4 + 0.01 + 0.89 + 0.37 = 8.53$. This value is also not statistically significant, p = .13, and this chi-square analysis also does not provide evidence for us to label this item as exhibiting DIF.

Mantel-Haenszel Odds Ratio

The Mantel-Haenszel approach is conceptually and computationally similar to the chi-square procedure. A set of 2 x 2 tables is also constructed similar to the chi-square procedure described above, then an index is calculated for each of the ability groups:

$$\alpha_i = \frac{a_i d_i}{b_i c_i} \text{ with } i = \text{ability group} \qquad (6)$$

An overall Mantel-Haenszel index (α_{MH}) is then computed as the average of all the indices for each ability group:

$$\alpha_{\text{MH}} = \frac{\sum_i \frac{a_i d_i}{N_i}}{\sum_i \frac{b_i c_i}{N_i}} \text{ where } N_i = \text{number of cases in } i^{\text{th}} \text{ ability group} \qquad (7)$$

Using the same examples(Tables 2& 3 for items 17&25 respectively), a_{MH} = 2.26 for item 17 and a_{MH} = 1.45 for item 25.

The overall Mantel-Haenszel index is usually transformed into a scale that centres at 0 (MH D-DIF = -2.35*ln*(α_{MH})), with distance from 0 indicating the degree of bias between the compared groups (Angoff, 1993). A value between -1.0 and 1.0is suggested by Dorans(1989)as indicative of DIF.For item 11, MH D-DIF = -2.35*ln*(2.26) = -1.92, putting it marginally over the -1.0 threshold. The result for item 25 is within the threshold, MH D-DIF = -0.87. Overall, the MH analysis results for items 17 and 25 do not indicate substantial DIF for these items, consistent with the chi-square results previously. Both the chi-square and Mantel-Haenszel procedures, while computationally more complex and tedious than the delta plot method, can be implemented in Excel.

Comparison of Results

The five different DIF analysis methods used in this paper did not produce consistent resultsfor the items used for illustration (Table 6). Nevertheless, results from some methods did agree with each other. The delta-plot, chi-square and logistic regression results did not provide evidence of DIF for all items, although the *p* values for items 17 and 25 are smaller than those of items 6 and 10. While the Mantel-Haenszel approach seems to suggest otherwise, and the results from the ICC method support some evidence that the groups perform differently on items 17 and 25, the magnitude of DIF does not appear to be large.Thus, while the methods may not be strictly consistent, they do not appear to provide contradictory results. With this in mind, it becomes important to realise that multiple approaches to DIF analysis will provide a more complete picture of how an item functions across groups.

The issue of practicality is also important to consider here. For example, the IRT approach is obviously more complex to understand and implement than the MH method, and requires more specialized software that might not be familiar to practitioners, but in exchange it is more effective for shorter tests and tests with lower reliability than the MH method (Zwick, 1990). The delta plot method is easy to understand and implement but there is concern that this method confounds item discrimination with differential functioning, thus making it less effective when applied to groups that are not well matched in ability (Angoff, 1982). Eventually, the need to balance statistical strengths with practical utility of DIF analysis methods will have to rest with the practitioner.

Implications

This paper discusses the utility of DIF analysis in test development and presents several DIF analysis methods that are practical even in classroom settings. As shown, DIF analysis methods differ in their ease of use and in their computational complexity, and each analysis has its own strengths and weaknesses depending on the situation that it is used. While indicators of DIF for particular items may show on some methods but not on others, using multiple approaches can provide a better perspective on how the items function. For example, the MH procedure indicated that item 10 is behaving differentially, but the other 4 methods indicate that there is no significant evidence of DIF (Table 6). It is hoped that by presenting these methods, the practitioner will realise that item analysis is an integral part of the assessment process and that poorly developed tests can have very significant implications – for all test takers, but more so for disadvantaged groups.The diversity of Australia's classrooms has direct educational implications for linguistic and cultural minorities among the student population, in particular those who are immigrant NESB students.

In comparison to the United States, where DIF analysis tends to focus on two group comparisons, (African-Americans vs Caucasians as one pair, and native English speakers vs English language learners as the other pair), there is no analogy in Australian classrooms for the US African-American/Caucasian group differences. Implications for DIF analysis can therefore focus on differences between ESB and NESB students. Practical and very important implications of DIF analysis could influence intervention strategies for English language learners or non-English background students who typically fall into two categories: English immersion and bilingual education (Slavin & Cheung, 2005). There is evidence that bilingual education approach might be more effective (e.g., Slavin& Cheung, 2005), although there are also some studies that offer more ambiguous results (e.g, Barnett, Yarosz, Thomas, Jung, & Blanco, 2007). What is more convincing, however, is the substantial evidence that language intervention offers significant effect sizes in terms of improvement for minority linguistic groups. Slavin and Cheung's metaanalysis found that bilingual education approaches are more effective in producing positive effect sizes compared to immersion approaches (Slavin & Cheung, 2005).

In Barnett and colleagues' (2007) study, both approaches seem to work equally well and both increased the performance of students whether they are English language learners or native English speakers. It appears that even if the choice of approach is less clear, there is a strong support that language intervention should have a positive impact on second language learners especially in academic areas where language skills may moderate the relationship between content learning and achievement. The preceding discussion on the implications of DIF on groups based on linguistic background is particularly relevant to Australia. As a more general framework, however, DIF on groups based on other variables has implications that are no less important – it would certainly be a worthwhile endeavour to investigate DIF between boys and girls, or between students of low and high SES. Regardless of what grouping variables are used, DIF analysis can help identify the factors behind group differences in test performance. If there are factors that are amenable to educational intervention, knowing what they are and quantifying their effects should translate into a more effective intervention program that can target specific student groups who would benefit the most, and one of the first steps in doing this involves the use of DIF analysis as an important tool in investigating test bias.

Methods to detect DIF in a classroom or school level will be increasingly useful in Australia as its school demographics become more diverse. The role of the government and federal or state agencies in minimizing the existence of DIF in national or state-wide tests will undoubtedly become more important, but we should also not disregard the role of individual schools in this matter. The readers and practitioners who might be in a role to implement DIF analysis methods and interpret the results in the future will have to balance the need for statistical rigor of a method with the practical considerations in terms of logistics and operational factors.

References

Angoff, W. H. (1972, September). A technique for the investigation of cultural differences. Paper presented at the annual meeting of the American Psychological Association, Honolulu, Hawaii.

Angoff, W. H. (1982). Use of difficulty and discrimination indices for detecting item bias. In R. Berk (Ed.), Handbook of Methods for Detecting Item Bias (pp. 96-116). Baltimore, MD: The Johns Hopkins University Press.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), Differential Item Functioning (pp. 3-23). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

Angoff, W. H., & Ford, S. F. (1974). Inter-race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 10, 95-105.

Australian Bureau of Statistics. (2006). A Picture of the Nation: The Statistician's Report on the 2006 Census. Retrieved from http://www.abs.gov.au/ausstats/subscriber.nsf/LookupAttach/2070.0Publication29.01.091/$File/20700_A_Picture_of_the_Nation.pdf.

Barnett, W. S., Yarosz, D. J., Thomas, J., Jung, K., & Blanco, D. (2007). Two-way and monolingual English immersion in preschool education: An experimental comparison. Early Childhood Research Quarterly, 22(3), 277-293.

Boehm, A. (2001). Boehm Test of Basic Concepts (3rd ed.). New York: The Psychological Corporation.

Bollen, K. A., & Long, J. S. (1993). Testing structural equation models. Sage focus editions, Vol, . 154. Thousand Oaks(1993).

Broughman, S. P., Swaim, N. L., & Keaton, P. W. (2009). Characteristics of Private Schools in the United States: Results From the 2007-08 Private School Universe Survey(NCES No. 2009-313). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Camilli, G. (1993). The case against item bias detection techniques based on internal criteria: Do item bias procedures obscure test fairness issues? In P. W. Holland & H. Wainer (Eds.), Differential Item Functioning (pp. 389-396). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

Care, E., Roberts, E., & Thomas, A. (2009). Effects of language background on measures of ability of children in their first year of school. Australian Educational and Developmental Psychologist, 26(1), 20-35.

Clauser, B. E., & Mazor, K. M. (1998). Using Statistical Procedures to Identify Differentially Functioning Test Items. Educational Measurement: Issues and Practice, 17(1), 31-44.

Department of Immigration and Citizenship. (2008). The People of Victoria: Statistics from the 2006 Census: Commonwealth of Australia.

Department of Justice. (2005). Research into health promotion and best practice services for culturally and linguistically diverse communities (Research report). Melbourne, VIC: Victorian Government Department of Justice.

Dorans, N. J. (1989). Two New Approaches to Assessing Differential Item Functioning: Standardization and the Mantel--Haenszel Method. Applied Measurement in Education, 2(3), 217.

Gipps, C., & Murphy, P. (1994). A fair test? Assessment, achievement and equity. Bristol, PA: Open University Press.

Hauger, J. B., & Sireci, S. G. (2008). Detecting differential item functioning across examinees tested in their dominant language and examinees tested in a second language. International Journal of Testing, 8, 237-250.

Howell, D. (2002). Statistical Methods for Psychology (5th ed.). Pacific Groove, CA: Duxbury.

Ironson, G. (1982). Use of chi-square and latent trait approaches for detecting item bias. In R. Berk (Ed.), Handbook of Methods for Detecting Item Bias (pp. 117-160). Baltimore, MD: The Johns Hopkins University Press.

Irtel, H. (1995). An extension of the concept of specific objectivity. Psychometrika, 60(1), 115-118.

Kelderman, H. (1988). Common item equating using the loglinear Rasch model. Journal of Educational Statistics, 13(4), 319-336.

McNamara, T., & Roever, C. (2006). Psychometric Approaches to Fairness: Bias and DIF. Language Learning, 56(Suppl 2), 81-128.

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

Ministerial Council for Education, Early Childhood Development, and Youth Affairs (MCEECDYA). (2008). National Assessment Program – Literacy and Numeracy (NAPLAN). Retrieved March 18, 2010, from http://www.naplan.edu.au/about/national_assessment_progralitеracy_and_numeracy.html

Penfield, R. D., Giacobbi, P. R., & Myers, N. D. (2007). Using the cumulative common log-odds ratio to identify differential item functioning of rating scale items in the exercise and sport sciences. Research Quarterly for Exercise and Sport, 78(5), 451-464.

Rasch, G. (1966). An item analysis that takes individual differences into account. British Journal of Mathematical and Statistica Psychology, 19, 49-57.

Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. Danish Yearbook of Philosophy, 14, 58-94.

Ryan, C., & Watson, L. (2004). The Drift to Private Schools in Australia: Understanding its Features (Discussion Paper No. 479): The Australian National University: Centre for Economic Policy Research.

Scheuneman, J. D. (1979). A method of assessing bias in test items. Journal of Educational Measurement, 16, 143-152.

Scholderer, J., Brunso, K., Bredahl, L., & Grunert, K. G. (2004). Cross-cultural validity of the food-related lifestyles instrument (FRL) within Western Europe. Appetite, 42(2), 197.

Shepard, L., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. Journal of Educational Statistics, 6(4), 317-375.

Slavin, R. E., & Cheung, A. (2005). A synthesis of research on language of reading instruction for English language learners. Review of Educational Research, 75(2), 247-284.

Swaminathan, H., & Rogers, H. J. (1990). Detecting Differential Item Functioning Using Logistic Regression Procedures. Journal of Educational Measurement, 27(4), 361-370.

Wilson, M., Allen, D. D., & Li, J. C. (2006). Improving measurement in health education and health behavior research using item response modeling: introducing item response modeling. Health Education Research, 21(Supplement 1), i4-i18.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). Conquest 2.0 Generalised Item Response Modelling Software (Manual). Camberwell, Victoria: Australian Council for Educational Research Ltd.

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

Zumbo, B. D., & Gelin, M. N. (2005). A matter of test bias in educational policy research: Bringing the context into picture by investigating sociological/community moderated (or mediated) test and item bias. Journal of Educational Research & Policy Studies, 5(1), 1-23.

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? Journal of Educational Statistics, 15(3), 185-197.

Table 1: Item Difficulty and Delta Values for Groups A and B

Item[a]	Group A *p*-values	Group B *p*-values	Group A deltas	Group B deltas	Distance from trendline[b]
item 1 (1)	0.95	0.89	1.01	1.21	-0.22
item 2 (2)	0.63	0.46	-0.91	-0.65	-0.13
item 3 (4)	0.90	0.60	0.74	-0.01	0.51
item 4 (5)	0.93	0.85	0.94	1.08	-0.17
item 5 (8)	0.66	0.46	-0.71	-0.65	0.00
item 6 (9)	0.74	0.54	-0.25	-0.28	0.04
item 7 (11)	0.89	0.79	0.68	0.81	-0.14
item 8 (12)	0.76	0.38	-0.12	-1.01	0.67
item 9 (14)	0.97	0.95	1.14	1.49	-0.34
item 10 (16)	0.76	0.54	-0.12	-0.28	0.13
item 11 (17)	0.59	0.56	-1.11	-0.19	-0.60
item 12 (19)	0.84	0.69	0.35	0.35	-0.03
item 13 (21)	0.96	0.84	1.07	1.03	-0.05
item 14 (23)	0.85	0.76	0.41	0.67	-0.22
item 15 (27)	0.92	0.92	0.88	1.35	-0.41
item 16 (29)	0.92	0.77	0.88	0.72	0.06
item 17 (31)	0.42	0.28	-2.17	-1.42	-0.40
item 18 (34)	0.89	0.72	0.68	0.49	0.09
item 19 (40)	0.91	0.79	0.81	0.81	-0.05
item 20 (41)	0.62	0.39	-0.98	-0.96	0.06
item 21 (42)	0.65	0.35	-0.78	-1.10	0.29
item 22 (45)	0.89	0.78	0.68	0.76	-0.11
item 2346)	0.37	0.11	-2.43	-2.14	-0.04
item 24 (47)	0.67	0.31	-0.65	-1.28	0.52
item 25 (49)	0.77	0.43	-0.05	-0.78	0.55

[a]Numbers in parenthesis are the original numbering of the items in BOEHM-3

[b]Positive values in the distance indicate advantage (i.e., item is easier) for group A, negative values indicate advantage for group B.

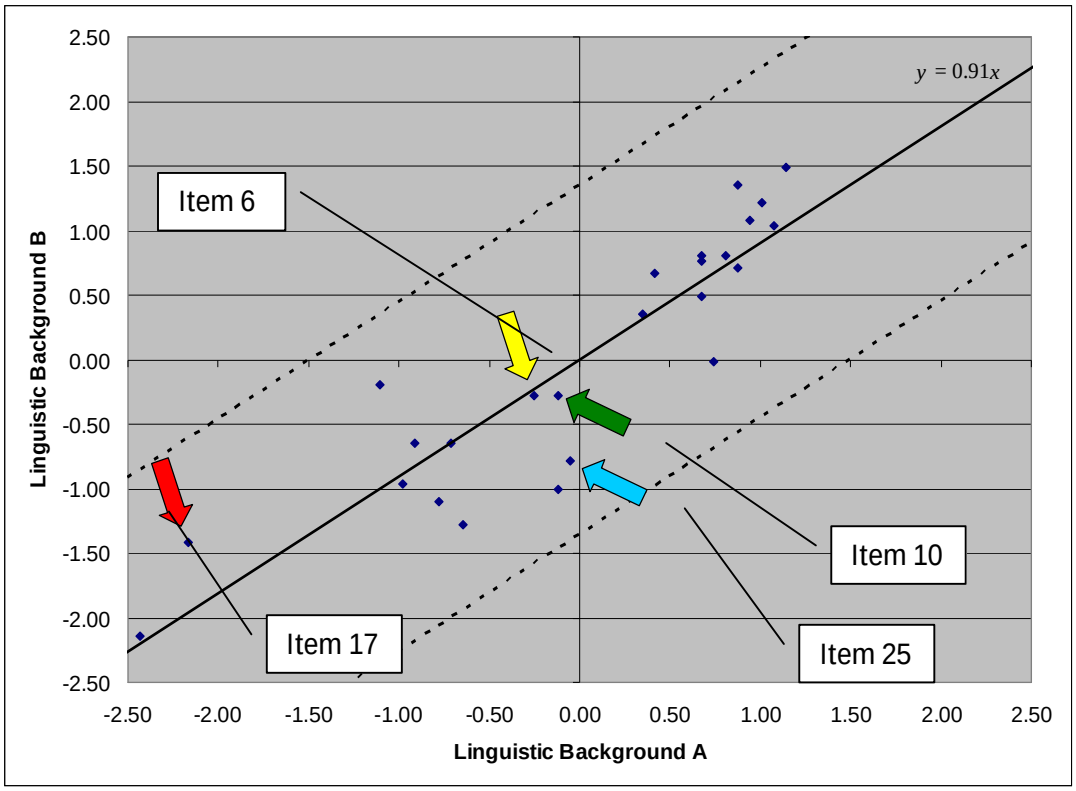

Figure 1. Delta plot of the 2 group data. Arrows for item 6, item 10, item 17, and item 25 indicate location relative to the trend line. Dashed line indicates a distance of -1.0 to 1.0 from the trend line

Table 2: Tables by Ability Level for Item 17

Ability level	Group	Number correct	Number incorrect
1	Group A	5 (*a*)	2 (*b*)
	Group B	17 (*c*)	10 (*d*)
2	Group A	6	0
	Group B	13	5
3	Group A	16	1
	Group B	26	4
4	Group A	24	3
	Group B	12	3
5	Group A	33	1
	Group B	6	0

Table 3: Tables by Ability Level for Item 25

Ability level	Group	Number correct	Number incorrect
1	Group A	0	7
	Group B	3	24
2	Group A	3	3
	Group B	1	17
3	Group A	6	11
	Group B	11	19
4	Group A	20	7
	Group B	9	6
5	Group A	32	2
	Group B	6	0

Table 4: Response Model Parameter Estimates Between Group a and Group B, with Groups Based on Linguistic Background

Variables		Estimate	Error	Difference (logits)[a]	Fit statistics (weighted)			
Item	Linguistic group				MNSQ	CI		*t*
6	A	0.08	0.17	-0.15	1.01	0.73	1.27	0.10
	B	-0.08	0.17		0.85	0.83	1.17	-1.90
10	A	0.00	0.18	0.00	1.01	0.71	1.29	0.10
	B	0.00	0.18		1.18	0.83	1.17	2.00
17	A	0.27	0.17	-0.54	0.99	0.82	1.18	-0.10
	B	-0.27	0.17		1.12	0.77	1.23	1.10
25	A	-0.35	0.18	0.69	0.95	0.69	1.31	-0.20
	B	0.35	0.18		1.00	0.83	1.17	0.00

[a]For the estimated difference, a positive value indicates advantage (i.e., item is easier) for group A, a negative value indicates advantage for group B.

Table 5: Logistic Regression Analysis Results for Items 6, 10, 17, and 25

Item	Factor	B	S.E.	Wald	p	exp(B)	95% CI for exp(B)	
							Lower	Upper
6	group	0.02	0.37	0.00	.95	1.02	0.50	2.10
	interaction	-0.05	0.09	0.28	.59	0.95	0.80	1.13
10	group	-0.40	1.37	0.08	.77	0.67	0.05	9.85
	interaction	-0.01	0.18	0.00	.97	0.99	0.69	1.42
17	group	0.58	0.54	1.17	.28	1.79	0.62	5.17
	interaction	0.04	0.10	0.15	.70	1.04	0.86	1.26
25	group	0.35	0.42	0.69	.41	1.42	0.62	3.27
	interaction	-0.06	0.13	0.23	.63	0.94	0.73	1.21

Table 6: Summary of Results for the Five DIF Analysis Methods

Item	Delta plot		IRT		Logistic regression		Chi-square method		Mantel-Haentzel	
	Distance	Indicator of DIF	Logits[a]	Indicator of DIF	Wald	p	χ^2	p	D-DIF	Indicator of DIF
6	0.04	none	-0.15	none	0.00	.95	1.16	.95	-0.36	none
10	0.13	none	0.00	none	0.08	.77	2.38	.79	1.04	minimal
17	-0.40	none	0.54	minimal	1.17	.28	3.72	.59	-1.92	some
25	0.55	none	-0.69	some	0.69	.41	8.53	.13	-0.87	minimal

[a]Difference in parameter estimates between groups

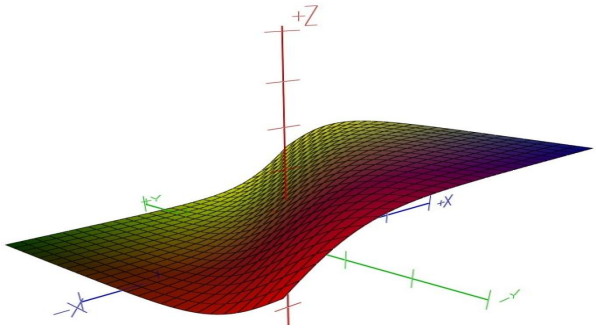

Figure 2°: Geometric Representation of P(X=1) in a Single-Parameter IRT Model

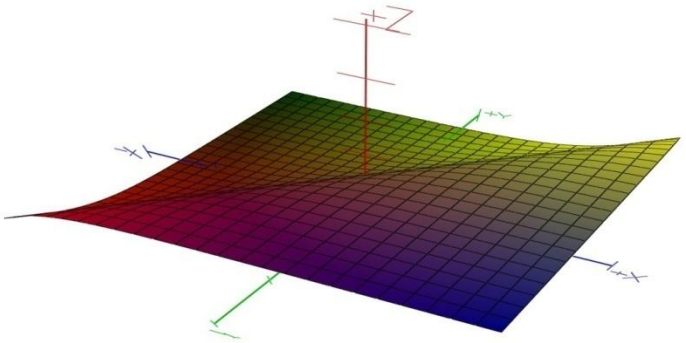

Figure 2b. Another view of Fig. 2a

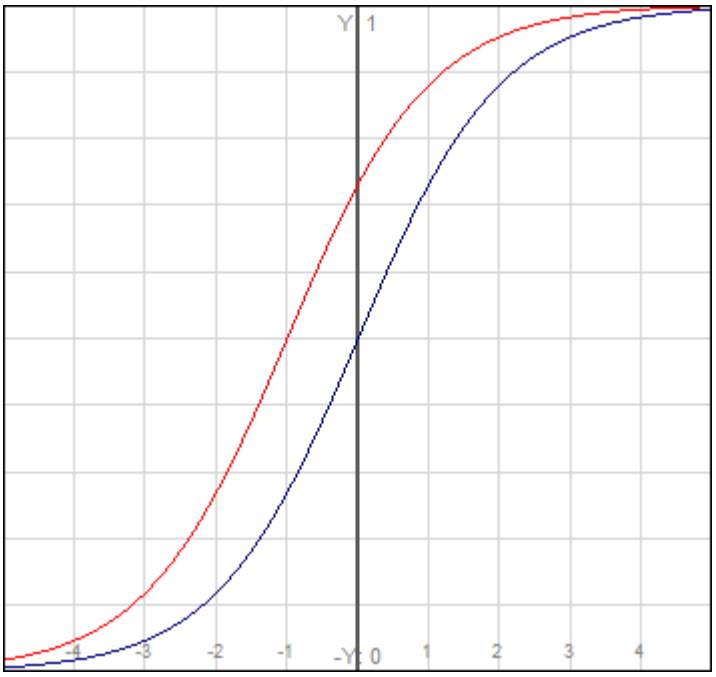

Figure 3: Item Characteristic Curves Showing How 2 Groups Perform Differently in an Item