Detecting Differential Item Functioning of Polytomous Items in Small Samples: Comparison of MIMIC with a Pure Anchor and MIMIC-Interaction Methods

Gavin T. L. Brown[1], Maryam Alqassab[1], Okan Bulut[2] & Jiaying Xiao[2]

[1] The University of Auckland
[2] University of Alberta

Abstract:
Differential item functioning (DIF) may be a result of either item bias or a real difference depending on whether the source of DIF is either construct-irrelevant or construct-relevant. It is relatively more challenging to conduct DIF studies when the sample size is small (i.e., < 200), items follow polytomous scoring (e.g., Likert scales) instead of dichotomous scoring, and psychological grouping variables are used instead of demographic grouping variables (e.g., gender). However, the multiple indicators-multiple causes (MIMIC) approach can be a promising solution to address the aforementioned challenges in DIF studies. This study aims to investigate the performance of two MIMIC methods, namely MIMIC with a pure anchor (MIMIC-PA) and MIMIC-interaction methods, for DIF detection in the Student Conceptions of Assessment inventory based on a psychological grouping variable derived from students' self-efficacy and subject interest. The results show that MIMIC-PA identified five mathematics and eight reading items with large DIF in the four factors. MIMIC-interaction showed that no items had uniform DIF, while four items had non-uniform DIF. Items with statistically significant DIF were aligned with the known effects of self-efficacy and subject interest on academic achievement, supporting the claim that observed DIF reflects item impact rather than bias. The study's implications for practice and directions for future research with the MIMIC approach are discussed.

Keywords:
Differential item functioning, polytomous scales, small sample, MIMIC

Detecting Differential Item Functioning of Polytomous Items in Small Samples: Comparison of MIMIC with a Pure Anchor and MIMIC-Interaction Methods

Within a framework of unidimensionality, test and survey items are assumed to measure a single and intended dimension of interest (e.g., reading performance in a reading assessment or motivation in a motivation-related scale) without being influenced by construct-irrelevant factors (e.g., ethnicity, race, or gender) that might contribute to differential performance among subgroups of the target examinee population. Differential item functioning (DIF) identifies, at the item level, whether a factor other than the intended dimension impacts upon the probabilities of answering items. If subgroups of an examinee population (e.g., boys vs. girls) have different probabilities of choosing an answer or response option, after being matched on a total score, then DIF is present (Zumbo, 1999), leading to false conclusions about the examinees' performance.

Detecting DIF is relatively straightforward in the context of dichotomously scored items (e.g., right or wrong) and demographic grouping variables (e.g., gender and ethnicity). However, when it comes to self-reports of psychological factors (e.g., beliefs, values, attitudes, or motivations) with polytomously-scored items, detecting DIF items becomes more challenging for several reasons. First, the underlying reason of DIF can be real-world differences in the psychological factors, instead of a demographic characteristic. For example, indigenous Māori students experience educational achievement quite differently compared to majority ethnicity students in New Zealand and this has been reflected in statistically significant differences in how their conceptions of assessment relate to academic achievement (Hirschfeld & Brown, 2009). In such a situation, it could be argued that DIF reflects the presence of a construct-relevant factor rather than bias from a construct-irrelevant factor (Zumbo, 1999, 2007). Previous research used DIF to detect strengths and weaknesses of students' subpopulations based on the assumption that students from different countries have different learning experiences (Klieme & Baumert, 2001).

Another challenge in detecting DIF in polytomously-scored items is sample size. DIF studies generally require large sample sizes (i.e., > 500) in both focus and reference groups to ensure accurate estimates of DIF (Zumbo & Witarsa, 2004). Surveys, rating scales, and similar instruments involving polytomous items require even large sample sizes in order to detect items exhibiting DIF across different response options. However, small sample size (i.e., < 200 in reference and focal groups) is a common problem in real-world education and psychology research, especially when self-report measures are used for data collection. Therefore, researchers who want to investigate item-level bias in polytomously-scored items from such measures need to choose a robust method that can detect items with significant DIF in small samples.

Previous studies indicated that the multiple indicators-multiple causes (MIMIC) methods can be a promising solution to detecting DIF in both dichotomous and polytomous items (e.g., Bulut & Suh, 2017; Lee, Bulut, & Suh, 2017; Shih & Wang, 2009; Wang, Shih, & Yang, 2009; Woods & Grimm, 2011). However, these studies either focused on the detection of DIF in dichotomous

items (e.g., Bulut & Suh, 2017; Lee, Bulut, & Suh, 2017; Finch, 2005) or used Monte Carlo simulations to generate large datasets with polytomous items (e.g., Wang & Shih, 2010). Therefore, little is known about the performance of MIMIC methods in detecting DIF in polytomous items, especially when the sample size is small. Therefore, this study aims to investigate the performance of two MIMIC methods in identifying polytomous items with DIF when the sample size is small and the grouping variable is a psychological variable instead of a demographic variable. Using real data from a self-report student inventory (Brown, 2008), the performance of MIMIC with a pure anchor (MIMIC-PA; Shih & Wang, 2009) and MIMIC-interaction (Woods & Grimm, 2011, Lee, Bulut, & Suh, 2017) methods in detecting DIF in polytomous items are compared. As the grouping variable, a combination of two psychological variables (self-efficacy and subject interest) is used, instead of a demographic variable. The results of this study will indicate whether the two MIMIC methods (i.e., MIMIC-PA and MIMIC-interaction) can identify polytomous items with DIF when the sample size is not very large. Furthermore, the results will also demonstrate whether the two MIMIC methods provide similar results with regard to the items flagged for exhibiting DIF.

Literature Review

Student Psychology around Assessment

How students perceive, feel, or think about assessment has become an important aspect of educational psychology (McMillan, 2016). Four major beliefs about the purposes of assessment have been identified (i.e., it is for improved learning and teaching; it is irrelevant; it creates positive emotions and classroom climate; it indicates external factors; Brown & Hirschfeld, 2008). Student achievement emotions have been classified as either pleasant or unpleasant and whether they activate further learning or not (Vogl & Pekrun, 2016). The emotional and motivational impact of assessment is most salient when students are given feedback as to their performance (Brown, Peterson, & Yao, 2016; Peterson, Brown, & Jun, 2015). McMillan (2016) has argued that the impact of assessment on the psychology of the learner begins before the assessment as the learner prepares, continues during the assessment processes proper, and culminates once the student knows how their performance was evaluated and as they begin to prepare for the next round of learning-assessment-feedback.

Research has shown that students who endorsed the notion that assessment supports greater learning had better performance, while endorsement of the conception that assessment can be ignored led to reduced performance (Brown, Peterson, & Irving, 2009). Wise and Cotten (2009) showed that students who endorsed improvement as a purpose of assessment guessed less than those who endorsed the conception that assessment was irrelevant. How students perceive novel forms of assessment (e.g., self-assessment, peer assessment, portfolio assessment, etc.) shapes their reaction to and engagement with these new evaluative methods (Struyven & Devesa, 2016). Students who experience activating emotions when assessed, regardless of whether those emotions are pleasant or unpleasant, tend to perform better (Vogl & Pekrun, 2016). Hence, there is evidence that student be-

liefs, emotions, and thoughts about the nature, purpose, and effect of assessment matter. It has been argued that how students manage their beliefs and feelings about assessment is a manifestation of self-regulated learning (Brown, 2011).

No single study can integrate or evaluate the many control and competence beliefs that have been found to predict and influence achievement (Schunk & Zimmerman, 2006). Hence, this study exploited two available predictors of achievement (i.e., self-efficacy and interest). An important competence belief is self-efficacy (i.e., "people's judgments of their capabilities to organize and execute courses of action required to attain designated types of performances" (Bandura, 1986, p. 391) because these beliefs influence the courses of action people choose and persist with, even in face of difficulties (Pajares, 1996). Importantly, self-efficacy is task and situation specific and is normally generated as a consequence of mastery experiences within specific domains (Bong, 2013). Self-efficacy as a predictor of performance in school subjects has standardised regression weights of between .20 and .55, suggesting weak to moderate amounts of variance are explained by self-efficacy, with notable variability by school subject or domain and by student overall ability (i.e., self-efficacy is more influential for lower-achieving students; Bong, 2013; Pajares, 1996).

An important control belief that influences the processes of selecting actions and outcomes is interest (Zimmerman & Schunk, 2004). Alexander (2003) distinguishes between individual (i.e., long-term enduring investment individuals have in a domain or facet of it) and situational (i.e., temporary arousal sparked by events or features of the 'here and now' environment) interest. The model of domain learning (MDL) proposes that early in the learning process, when student knowledge or competence is relatively low, situational interest is useful in motivating students to learn (Alexander, 1995). As knowledge competence grows, individual interest develops, sustaining motivation to learn an increasing complex and sophisticated understanding of the domain.

Thus, both self-efficacy in a specific domain and interest in the same domain have been shown to be statistically significant predictors of performance. A large-scale survey of New Zealand students in Grades 5 to 12 ('Otunuku & Brown, 2007) has shown that there are moderate inter-correlations between self-efficacy and interest in reading comprehension (r=.64), writing (r=.72), and mathematics (r=.65). The same study reported that the inter-correlations of interest and self-efficacy with standardised test scores in reading, writing, and mathematics had small positive associations. Hence, at least among New Zealand students it has been shown that conceptions of assessment, self-efficacy, and interest in the subject all positively regress onto achievement.

In light of these findings, the study presented here has examined the DIF properties of student conceptions of assessment for four groupings of students (i.e., high vs. low interest and self-efficacy by two school subjects). Given that higher interest and self-efficacy is associated with greater performance, as is greater endorsement of assessment improves teaching and learning, it could be expected that DIF may appear among conceptions of assessment items that are conceptually aligned with interest or self-efficacy. If this is the case then it could be that those items reflect real-world differences in the groups rather than some deficiency of measurement.

Current Study

This study examines the Students Conceptions of Assessment (SCoA version 6) inventory items (Brown, 2008) for DIF using participant groups defined by whether they were high versus low in self-efficacy and interest in two different school subjects (i.e., mathematics and reading). Two techniques of DIF analysis (i.e., MIMIC-PA and MIMIC-interaction) were applied for detecting self-report SCoA items with uniform and non-uniform DIF. Within each subject, DIF analysis with the MIMIC-PA method was separately conducted for each of the four SCoA factors (i.e., improvement, externality, affect, and irrelevance), while all four factors were analysed simultaneously using the multidimensional form of the MIMIC-interaction method (Lee et al., 2017). As the grouping factors, students' levels of interest and self-efficacy were used. DIF analyses were completed separately for each school subject (i.e., math or reading) to better identify subject-specific effects. Results from the MIMIC-PA and MIMIC-interaction methods were compared and the implications for practice were discussed.

Methods

Sample

The data for this study had been previously collected from complete cohorts of students in the first two years of high school (Grades 9-10, nominal ages 13-14) from a small number of schools in the Auckland metropolitan region (results reported in: Brown, Irving, Peterson, & Hirschfeld, 2009; Brown, Peterson, & Irving, 2009). The data file for this study is available on the first author's institutional data repository[1]. Voluntary participation in the study was obtained with informed written consent according to the University of Auckland Human Participant Ethics Committee guidelines (ref: 2004/456).

A total of 803 valid cases (i.e., 411 in reading and 392 in mathematics) remained after removing students who did not have scores for the standardised achievement test or for self-efficacy and interest (see Table 1). The sample was almost equally split between sexes, with ages ranging from 13 to 16 years. Consistent with New Zealand 2006 Census data for children age 5 to 19, the ethnic mix and sex distribution of the sample was statistically equivalent to the population of school-age children. Accordingly, the sample used in this study is representative of the demographics of the New Zealand high school student population.

The grouping variable was the sum score of the three self-efficacy and three interest items. To maximize the sample size based on the grouping variable (i.e., the sum of subject-specific self-efficacy and interest), the composite scores were split at the mean in order to create two groups, each of which would be as close as possible to the target sample size of 200. Students whose scores fell exactly on the mean were discarded and the rest were classified as having either high or low levels of self-efficacy or interest. The four groups were unequal and all were close to but below 200 (Reading: High n = 188, Low n = 184; Mathematics: High n = 189, Low n = 199). A factor that makes analyzing these data challenging is that the sample sizes are below the recommended threshold of $n > 200$ in focal and reference groups (Zumbo, 1999).

1 doi:10.17608/k6.auckland.7688651

Table 1 Demographic Characteristics of the Participants

Demographic Trait		% or Mean (SD)	Difference to NZ population
Sex			χ^2=0.08, p=.78
	Male	49%	
	Female	51%	
Age		13.92 (.72)	
Ethnicity			χ^2=7.59, p=.11
	NZ European	48 %	
	Māori	12%	
	Asian	18%	
	Pasifika	8%	
	Other	14%	

Note Māori=indigenous people of New Zealand; Pasifika=immigrant peoples from Pacific Island nations.

Instruments

Students' Conceptions of Assessment (SCoA-VI) inventory

The SCoA inventory is a multidimensional self-report survey. Previous studies found that the SCoA inventory met full measurement invariance conditions for sex, ethnicity, and year groups (Hirschfeld & Brown, 2009), whereas only metric equivalence was found when students had different levels of interest and self-efficacy in reading (Brown & Walton, 2017). The absence of full measurement invariance for some psychological factors, in contrast to demographic factors, raises the possibility that, if there are items with DIF in the SCoA inventory, it could be due to a construct-relevant factor rather than bias.

The SCoA-VI inventory consists of 33 self-report items to identify four inter-correlated assessment-related beliefs: namely, improvement, externality, affect, and irrelevance (Brown, 2008). Improvement involves both the teacher and the student using assessment to improve learning. Externality in the SCoA inventory refers to attributing the consequences of assessment to external uncontrollable sources such as teachers and schools or one's future career or IQ. The affect factor focuses on the emotional and social impacts of assessment on students. Perceiving assessment as bad, interfering with learning, or choosing to ignore it are captured by the irrelevance factor. Participants respond using a six-point positively-packed (Lam & Klockars, 1982) rating scale with four degrees of agreement (6 = Strongly Agree, 5 = Mostly Agree, 4 = Moderately Agree, 3 = Slightly Agree) and two negative categories (2 = Mostly Disagree, 1 = Strongly Disagree). When participants respond to socially desirable statements, they are more likely to agree, which reduces variability and precision in the data; hence, positive packing (i.e., giving more positive than negative choices in the response scale) has been argued as appropriate in these circumstances.

The overall fit of the model with two samples has been good; for example, Brown, Irving, Peterson, and Hirschfeld (2009) reported a sample of 705 had good model-da-

ta fit; $\chi^2_{(481)}$ = 1551.57, χ^2/df = 3.23, p = 0.07, CFI = 0.89, GFI = 0.92, RMSEA = 0.056, 90% CI = 0.053-0.059, SRMR = 0.060. For the data used in this study, both good fit (N_{2006} = 705; N_{2007} = 624; number of manifest variables = 66; χ^2 = 3104.85; df = 960; χ^2/df = 3.234 (p = .07); CFI = .89; GFI = .95; RMSEA = .041, 90% CI = .039–.043; SRMR = .060) and strong invariance (i.e., equivalent regressions and intercepts) was demonstrated to the earlier data (Brown, Peterson, & Irving, 2009). The internal estimate of reliability (McDonald's ω) for each of the four SCoA factors of interest in this dataset were good (i.e., ω_{Affect} =.90; $\omega_{\text{Improvement}}$ =.89; $\omega_{\text{Irrelevance}}$ =.83; $\omega_{\text{Externality}}$ =.77). Studies with this version of the SCoA have found statistically significant regressions to standardized test scores (i.e., performance). For example, in a national survey of 31 high schools in New Zealand, the improvement conception predicted higher mathematics scores (β = .65), while endorsement of the externality conception predicted lower mathematics scores (β = -.82; Brown, Peterson, & Irving, 2009).

Student self-efficacy and interest

Six items are administered as part of a school administered, standardized achievement test from the Assessment Tools for Teaching and Learning (asTTle) system (Hattie et al., 2005) distributed to schools by the New Zealand Ministry of Education. Participants indicate their agreement by selecting one of four smiley-face options (i.e., 1 = ☹, 2 = ☹, 3 = ☺, and 4 = ☺). Three items related to self-efficacy in the subject being tested (reading or mathematics) and three related to interest in the subject. Confirmatory factor analysis of these items based on the norming data showed that two factors underlay the items (Reading: n = 29337, χ^2 = 524.67, df = 8, CFI = .99, GFI = .99, RMSEA = .047; Mathematics: n = 22413, χ^2 = 610.07, df = 8, CFI = .99, GFI = .99, RMSEA = .058; 'Otunuku & Brown, 2007). While the correlation between the two factors is moderate (Reading r = .64; Mathematics r = .65), the asTTle system averages the responses to these six items and reports a total score of self-beliefs within the subject ranging from 1.00 to 4.00. It is not possible to disentangle this score and so the aggregate attitude to subject score is used as a continuous grouping variable.

Data Analysis

It is worth noting that while survey items are frequently evaluated for equivalence in factor analytic traditions through invariance testing (Tran, 2009), DIF analysis utilizing item response theory (IRT) is a parallel technique valid for items that use ordinal categorical responding (Grimm & Widaman, 2012). The integration of factor analytic and IRT methods for evaluating DIF or measurement invariance in ordinal categorical responses has become commonplace since the identification of appropriate techniques (Muthén, 1984). The modelling approach reported here is situated within the MIMIC framework (Jöreskog & Goldberger, 1975), in that multiple manifest variables ($y_1, \ldots, y_k$) are indicators of a latent trait (θ in IRT or y^* in factor analysis) and a group membership variable (z) giving the impact of psychological or demographic factors upon the latent trait. Thus, in a MIMIC model it is possible to determine the degree to which a latent trait is dependent on group membership.

It seems that the MIMIC approach works well with focal group sizes as small as 50 to 100, provided the reference group was large (i.e., n > 500) (Woods, 2009). Nonetheless,

the challenge for the MIMIC approach exists when both reference and focal groups are small to marginal (i.e., $n \leq 200$). This paper contributes to our understanding of DIF by comparing two MIMIC-based DIF detection methods (i.e., MIMIC model with a pure anchor [MIMIC-PA] and MIMIC-interaction model) and applying them to ordinal rating data in which students' psychological beliefs act as the grouping variable, allowing the possibility that DIF reflects a construct-relevant factor rather than bias. We have selected these different methods because both are appropriate for response data with small sample sizes.

It should be noted that despite sharing the MIMIC framework, the two DIF detection methods differ with regard to their implementation. The MIMIC-PA method requires an iterative purification process in which at least one item must be identified as an anchor item (i.e., DIF-free item) and the remaining items can be individually tested for DIF in the first model. In the subsequent models, items that do not exhibit DIF in the first model can also be used as anchor items. This process continues until the model identifies the same items that consistently exhibit DIF. To date, the MIMIC-PA model has only been used for detecting uniform DIF in dichotomous and polytomous items in unidimensional instruments (e.g., Shih & Wang, 2009; Wang & Shih, 2010).

As for the MIMIC-PA method, the MIMIC-interaction method also requires a set of anchor items to be identified before running DIF analysis. However, unlike the MIMIC-PA method, the MIMIC-interaction method does not necessarily require an iterative purification process for removing DIF-free items from the model. Instead, the model can be separately run for each item or a group of items under investigation. Furthermore, the MIMIC-interaction model is capable of testing both uniform and nonuniform DIF within the same model, using an interaction term between the latent trait and the grouping variable. To date, the MIMIC-interaction model has been applied to both unidimensional (Woods and Grimm, 2011) and multidimensional (Bulut & Suh, 2017; Lee et al., 2017; Woods & Grimm, 2011) instruments consisting of dichotomous items. However, the performance of the MIMIC-interaction method for detecting DIF in polytomous items is still unknown.

DIF detection with MIMIC

Self-reported surveys are generally based on the latent trait theory in which multiple items either indicate latent traits or influence those latent traits. An advantage of the MIMIC approach is that it can handle both dichotomous and polytomous item responses in the context of a factor analytic model (Zumbo, 1999). In ordinal rating scales, DIF occurs when the probability of responding to different categories of the rating scale differs for participants from different subgroups, after taking their levels in the target latent trait into account. Just as DIF studies are restricted to total scores for a unidimensional test measuring a single latent trait, in a multidimensional inventory it is conventional to analyze each dimension (i.e., latent trait) separately (Hamilton, 1999). However, the MIMIC approach is capable of analyzing DIF across multiple dimensions within the same model (Lee et al., 2017).

To counter over-sensitivity in detecting DIF with large samples, Holland and Thayer (1988) recommended determining the magnitude or effect size of DIF to differentiate between potential and minor DIF, which can be determined by log-odds ratios, beta-coefficient, and R^2 measures (Teresi et al., 2008).

For a dichotomous item, DIF can be constant or uniform when the probability of answering an item correctly is consistently greater for one group over all ability levels, or non-uniform when the difference in the probabilities of a correct response varies across ability levels. However, in the context of the MIMIC approach, the uniform and non-uniform DIF can be defined as a significant dependency between an item and an external variable (i.e., a grouping variable). Two MIMIC methods (MIMIC-PA and MIMIC-interaction) have recently entered the literature and are used in this study with real data. The following section describes how these two MIMIC methods work for detecting DIF.

MIMIC-PA

In the process of DIF detection, most approaches assume that all of the items that function as manifest indicators of the latent trait are unbiased except for the item being evaluated (Teresi, 2006). Holland and Thayer (1988) suggested that using a set of anchor items and purifying the latent trait of biased items can eliminate the contamination of the latent trait and thereby detecting DIF more accurately. A model for polytomous DIF detection based on a purified anchor design reduced Type I errors in over-identifying items with DIF (Shih & Wang, 2009; Wang & Shih, 2010). The MIMIC-PA approach identifies the average level of DIF in each item while iteratively holding each item as a DIF-free anchor so that the item with the smallest mean value acts as the anchor for the DIF analysis of other items. In a relatively small scale (e.g., less than 10 items), it is recommended that only one item is selected as the anchor.

In MIMIC-PA, the latent trait variable θ underlying item responses predicts the latent response variable (y_i^*)[2] (Wang & Shih, 2012, pp. 168-169). That factor, in accordance with MIMIC models, is simultaneously influenced by a grouping variable z. The ordinal options in the item response scale are the factor loadings (λ_i) from θ and relate to the IRT discrimination parameter of each item. The error term (ε_i) has a normal distribution for the ordinal probit and a logistic distribution for the ordinal logit. β_i is the effect of the grouping variable z on y_i^*. For item *i*, uniform DIF exists if there is a direct effect of z on y_i^*:

$$y_i^* = \lambda_i \theta + \beta_i z + \varepsilon_i \quad (1)$$

The left panel of Figure 1 provides a schematic framework of the MIMIC-PA approach; the latent trait θ predicts responses on self-reported items y1-y3, while the categorical variable z is the grouping variable that influences both the latent factor and items y_1 and y_2. There is no path from z to y_3 since that item has been set as the anchor variable. An item is identified as potentially having DIF if the statistical and practical significance of the DIF is established. The beta coefficient (β) for the item has to be statistically significant (e.g., $p < .05$) according to the Wald Test and the effect size (i.e., the item's beta coefficient divided by the item's loading) has to be notable (Shih & Wang, 2009).

MIMIC-interaction method

Unlike traditional MIMIC models, the MIMIC-interaction model (Woods & Grimm, 2011) contains an interaction term for the grouping variable and the latent trait variable that enables the model to detect both uniform and non-uniform DIF simultaneously. Previous

2 For the sake of brevity, we do not use a subscript for individuals (i.e., respondents) when explaining the formulation of the MIMIC models.

studies have shown that the MIMIC-interaction model was a useful tool to detect DIF in the context of both unidimensional and multidimensional assessments (Bulut & Suh, 2017; Lee et al., 2017; Woods & Grimm, 2011). In this study, a multidimensional form of the MIMIC-interaction model (Lee et al., 2017) was applied to the polytomously-scored SCoA items. That is, the SCoA items related to four constructs (i.e., improvement, externality, affect, and irrelevance) were tested for DIF within the same model.

For each polytomous item, the latent continuous response variable y_i^* could be represented as an observed ordinal response y_i via a threshold model, identical to that used in MIMIC-PA (Wang & Shih, 2012):

$$y_i = \begin{cases} 0 & \text{if } y_i^* \leq \tau_{i1} \\ 1 & \text{if } \tau_{i1} < y_i^* \leq \tau_{i2} \\ & \vdots \\ J & \text{if } y_i^* > \tau_{iJ} \end{cases}, \quad (2)$$

where τ_{ij} is the threshold parameter of step j (j = 1, 2, ..., J) in item i. Based on the MIMIC-interaction model, Equation 3 demonstrates y_i^* can be predicted by the latent trait θ, group variable z and their interaction term θz:

$$y_i^* = \lambda_i \theta + \beta_i z + \omega_i \theta z + \varepsilon_i \quad (3)$$

where λ_i is the factor loading of item i, β_i is the group difference in the threshold parameter after controlling for any mean ability difference on θ between groups. If $\beta_i \neq 0$ it indicates uniform DIF exists in item i. ω_i refers to the nonuniform DIF effects (when $\omega_i \neq 0$), and ε_i is the error term that is normally distributed and independent of θ and z. The multidimensional form of the model in Equation 3 involves θ and the interaction term of θz for each construct. The right panel of Figure 1 demonstrates the MIMIC-interaction model for detecting DIF in a polytomous item.

When compared to a MIMIC model, the MIMIC-interaction model seemed to exhibit higher false positive rates, especially when the anchor test is short and the magnitude of DIF is small (Lee et al., 2017; Woods & Grimm, 2011). The MIMIC-PA was also reported to maintain acceptable Type I error rates only with few DIF items in the test (Shih & Wang, 2009). It is, however, unclear which one of these two MIMIC based models is more likely to have higher inflation of Type 1 error rate using real-world data.

Results

Descriptive statistics for the asTTle attitude to subject, standardised achievement test score (aRs and aMs), and the four Student Conceptions of Assessment factors are provided in Table 2.

MIMIC-PA Analysis

Table 3 shows the DIF values for items except for the one used as the anchor item for the MIMIC-PA analysis of each SCoA factor. Both Irrelevance and Improvement SCoA factors shared the same item as an anchor across both school subjects. However, for SCoA Affective-Social and the External Attributions factors, the anchor item was different for reading and mathematics.

Results of the MIMIC-PA analysis identified five items with large DIF (i.e., ≥ ± 0.088; Gierl, Bisanz, Bisanz, Boughton, & Khaliq, 2001) in mathematics and eight items with large DIF in reading. In the SCoA Affective-Social factor item pe2 had large DIF favoring students with high levels of self-ef-

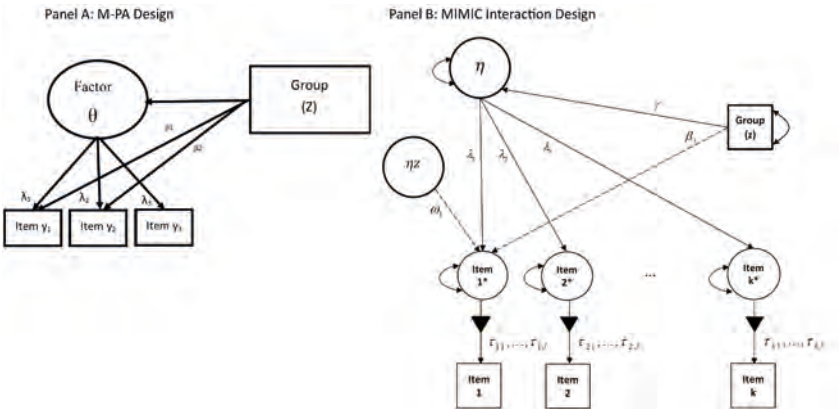

Figure 1 Schematic framework of the MIMIC-PA and MIMIC-Interaction models

Table 2

Measure	Reading M (SD) [N=411]	Mathematics M (SD) [N=388]
Attitude to Subject		
High Group	3.41 (.31)	3.27 (.34)
Low Group	2.43 (.37)	2.36 (.41)
Conceptions of Assessment		
Affective-Social	3.16 (1.13)	3.03 (1.09)
External	3.95 (1.00)	3.87 (0.96)
Improvement	4.36 (0.89)	4.34 (0.86)
Irrelevance	2.65 (0.95)	2.62 (0.94)

ficacy and interest in both mathematics and reading, while item ce6 favoured high levels of self-efficacy and interest in reading only. In SCoA External Attribution, no items showed any substantial DIF in mathematics; whereas, in reading, items sq1 and sq2 had large DIF favouring students who were both highly interested and self-efficacious. In SCoA Improvement, item ti6 had large DIF favoring highly motivated students in both reading and mathematics; while in reading item ti1 also had large DIF. In SCoA Irrelevance, item ig1 had large DIF for both school subjects in favour of highly motivated students. In mathematics, two further (bd1 and ig2) had large DIF both favouring low self-efficacy and interest students. In reading, item ig3 had large DIF favouring students who were highly motivated.

Table 3 MIMIC-PA DIF Results for the SCoA Items by Factor and Subject

	Mathematics		Reading	
Factor and Items	**DIF magnitude**	**Wald test**	**DIF magnitude**	**Wald test**
Affective-Social				
pe1: I find myself really enjoying learning when I am assessed	-0.093	-1.587	—	
pe2: Assessment is an engaging and enjoyable experience for me	0.139*	2.455	0.188**	3.534
ce1: When we do assessments, there is a good atmosphere in our class	0.046	0.818	-0.082	-1.335
ce2: When we are assessed, our class becomes more motivated to learn	—		-0.06	-1.047
ce3: Assessment motivates me and my classmates to help each other	-0.084	-1.512	-0.068	-1.169
ce4: Our class becomes more supportive when we are assessed	-0.06	-0.969	0.014	0.256
ce5: Assessment makes our class cooperate more with each other	-0.058	-1.137	-0.092	-1.634
ce6: Assessment encourages my class to work together and help each other	0.075	1.269	0.114*	2.014
External Attributions				
sq1: Assessment measures the worth or quality of schools	0.062	0.725	0.165**	4.579
sq2: Assessment provides information on how well schools are doing	0.048	0.52	0.14**	3.83
sf1: Assessment is important for my future career or job	-0.021	-0.214	0.046	1.242
sf2: Assessment results predict my future performance	-0.069	-0.568	—	
sf3: Assessment results show how intelligent I am	—		-0.032	-0.791
sf4: Assessment tells my parents how much I've learnt	0.036	0.633	-0.005	-0.138
Improvement				
si1: I look at what I got wrong or did poorly on to guide what I should learn next	0.091	0.507	0.16	1.665
si2: I pay attention to my assessment results in order to focus on what I could do better next time	—		—	
si3: I make use of the feedback I get to improve my learning	0.155	1.218	0.1	1.22
si4: I use assessments to take responsibility for my next learning steps	0.18	1.488	0.13	1.824
si5: I use assessments to identify what I need to study next	0.081	0.518	0.11	1.225
ti1: My teachers use assessment to help me improve	0.159	0.969	0.209*	2.382
ti2: Teachers use my assessment results to see what they need to teach me next	0.095	0.516	0.173	1.638

Factor and Items	Mathematics		Reading	
	DIF magnitude	Wald test	DIF magnitude	Wald test
ti3: Assessment shows whether I can analyse and think critically about a topic	-0.02	-0.103	0.085	0.829
ti4: Assessment is checking off my progress against achievement objectives or standards	0.155	0.916	0.079	0.832
ti5:Assessment is a way to determine how much I have learned from teaching	0.007	0.122	-0.03	-0.665
ti6: Assessment helps teachers track my progress	0.288*	5.085	0.223**	4.293
Irrelevance				
bd1: Assessment is value-less	-0.116*	-2.907	-0.066	-1.613
bd2: Assessment is unfair to students	-0.052	-1.446	0.051	1.252
bd3: Assessment results are not very accurate	—		—	
bd4: Teachers are over-assessing	-0.05	-1.292	-0.033	-0.741
bd5: Assessment interferes with my learning	-0.004	-0.108	0.034	0.74
ig1: Assessment has little impact on my learning	0.195**	3.701	0.199**	4.45
ig2: I ignore or throw away my assessment results	-0.109*	-2.479	-0.064	-1.523
ig3: I ignore assessment information	0.044	0.982	0.09*	2.111

Note — = pure anchor item; items with potential DIF indicated as $^{*}p < .05$, $^{**}p < .001$

MIMIC-Interaction Model Analysis

In contradiction to the MIMIC-PA approach, the MIMIC-interaction model (Table 4) showed no items had uniform DIF in either reading or mathematics. However, in the SCoA Improvement factor, the item si2 had nonuniform DIF in reading, while two items (ti4 and ti3) had nonuniform DIF in mathematics. In the SCoA Irrelevance factor, the item ig2 had nonuniform DIF in mathematics. Simply, fewer items with DIF were identified using the MIMIC-interaction model.

Table 4 MIMIC-Interaction Model DIF Results for the SCoA Items by Factor and Subject

	Mathematics		Reading	
Items	Uniform DIF	Nonuniform DIF	Uniform DIF	Nonuniform DIF
Affective-Social				
ce1	1.472	1.571	1.385	1.058
ce2	1.446	0.695	1.325	1.402
ce3	1.382	-0.014	1.316	1.251
ce4	1.328	-0.527	1.268	0.661
ce5	1.439	-0.227	1.283	0.856
ce6	1.334	0.225	1.338	0.404
pe1	1.637	-1.398	1.456	-0.955
pe2	1.593	-1.541	1.452	-1.653
External Attributions				
sf1	0.5	-0.561	-1.418	-0.569
sf2	0.434	-0.145	-1.57	-0.402
sf3	0.525	-0.261	-1.471	-0.068
sf4	0.496	-1.554	-1.305	0.804
sq1	0.45	0.361	-1.52	1.427
sq2	0.405	-0.054	-1.404	0.754
Improvement				
si1	-1.174	0.228	-1.335	-0.61
si2	-1.064	1.28	-0.831	**1.997***
si3	-1.186	1.21	-1.494	-0.053
si4	-1.075	1.696	-1.32	-0.317
si5	-1.187	1.525	-1.5	0.597
ti1	-1.176	0.183	-1.64	-0.149
ti2	-1.339	-0.158	-1.684	0.986
ti3	-1.095	**2.568***	-1.312	0.538
ti4	-1.144	**2.86****	-1.155	1.73
ti5	-1.045	0.968	-1.431	1.769
ti6	-1.095	1.267	-1.088	1.626
Irrelevance				
bd1	0.959	1.914	0.447	0.869
bd2	0.678	0.768	0.543	1.094
bd3	0.959	-0.905	0.841	-0.018
bd4	0.569	0.178	0.566	-0.675
bd5	0.722	-0.104	0.572	-0.945
ig1	0.569	0.715	0.554	0.49
ig2	0.239	**2.063***	0.344	0.57
ig3	0.541	0.961	0.452	1.201

Note Items with potential DIF are bolded. $^{*}p < .05$, $^{**}p < .01$

Discussion

Before discussing the substantive implications of this study, some consideration of the technical issues is worthwhile.

Small Sample DIF Analysis

While IRT-based DIF detection methods can handle both dichotomous and polytomous item responses as well as MIMIC, an advantage of MIMIC that is not shared by IRT-based DIF detection methods is the ability to include multiple "studied" background variables of different forms. Hence, both DIF methods deployed in this paper take advantage of having a psychological factor as the grouping condition in terms of explaining manifest and latent responses in the SCoA.

For the most part, the MIMIC-PA method reported relatively low levels of DIF, with just five items in mathematics and eight items in reading having large amounts of DIF, all purportedly uniform DIF. Indeed, a feature of the purification method is that, by treating one item as not having DIF, this process should result in fewer DIF items being identified. It is worth considering that in conducting the DIF analysis iteratively (i.e., finding the anchor with least DIF and then test for DIF) the effect of multiple testing has not been corrected. In other words, more items with DIF are detected potentially because more DIF tests were run.

In contrast, the MIMIC-interaction approach showed that only non-uniform DIF existed with just one item in reading and three items in mathematics. This suggests that, despite the technical complexity of the MIMIC-interaction model and the high false-positive rates associated with this model (Bulut & Suh, 2017; Lee et al., 2017; Woods & Grimm, 2011), it is much more effective in reducing instances of DIF Type I errors compared to the MIMIC-PA model. This method appears to be more robust than the MIMIC-PA model with small sample sizes. In addition, the multidimensional version of the MIMIC-interaction model takes the correlation among multiple latent traits into account and thereby improving the reliability of the estimation process and providing more accurate results.

Yet, the performance of both models could be further compared to investigate whether the MIMIC-interaction approach outperforms the MIMIC-PA model under different conditions (e.g., few DIF items, larger sample sizes, larger separation between groups). Additionally, artificial DIF (i.e., DIF favouring one group on some items leading to DIF favouring the other group on other items) can be misidentified as real DIF (Hagquist & Andrich, 2015). Thus, it is worthwhile to test in future analyses how much artificial DIF might be produced by the MIMIC-interaction and the MIMIC-PA models. Because there was no gold-standard for determining the true DIF in these real-world data, using another DIF detection approach (e.g., IRT likelihood ratio; Choi, Gibbons, & Crane, 2011) may have been able to corroborate the results reported here. This could be investigated in a future simulation study.

Unlike most DIF studies that use a dichotomous demographic variable, this study used a non-discrete psychological background variable (i.e., self-efficacy and interest toward school subjects) as the grouping variable. It is worth noting that while the grouping variable has a binomial distribution, this does not impact the estimation of the interaction term which is handled appropriately in Mplus. A more sophisticated way of classifying students into different levels of self-efficacy and interest is needed

in future studies, rather than simply splitting at the mean of the sample available. DeMars (2015), proposed a Mantel-Haenszel DIF detection approach based on a continuous rather than a categorical grouping variable which should be more suitable for use with latent traits (e.g., self-efficacy and interest) measured as continuous constructs. Future studies should consider if this approach could be extended to the MIMIC DIF models.

A strength of this paper is that it shows that with relatively small sample sizes and polytomous items, these two DIF detection methods differ significantly despite sharing the same MIMIC framework. There are multiple reasons for these different results. The MIMIC-PA is a unidimensional approach, is iterative in its search for DIF, and does not test for non-uniform DIF. In contrast, MIMIC-interaction not only tests both uniform and non-uniform DIF but is also multidimensional. These characteristics suggest it should be a preferred method for DIF detection in MIMIC contexts.

DIF in the SCoA Inventory

The MIMIC-interaction DIF analysis identified just four items altogether that had statistically significant non-uniform DIF. This interaction is more difficult to explain than uniform or consistent DIF as seen in the MIMIC-PA analysis. Here the advantage in these four items varies according to where the student is on the continuum for the item. Two of the items are within the teacher improvement aspect of the Improvement factor. It is possible, given the data were from a small number of schools, that there are real world differences in how teachers use assessments to diagnose and improve student competence. If this could be validated, then the observed DIF might reflect real-world conditions rather than bias. The non-uniform DIF for item si2 suggests that paying attention to assessments to guide next learning is unusually related to self-efficacy and interest in reading. The behaviours underlying ig2 (i.e., throwing away results) could be present with both high and lower self-efficacy and interest students, depending on the value they give to the test or the result obtained among other factors. Independent of their self-efficacy and interest related to the subject, both groups could endorse this response for quite different reasons (e.g., high students do not need results to validate their confidence or interest, while lower self-efficacy and interest students may disregard results because those are poor). The absence of uniform DIF may explain why an earlier study (i.e., Brown, Peterson, & Irving, 2009) reported strong invariance.

Thus, a tentative case could be made that these items, insofar as self-efficacy and interest are concerned, either do not have DIF at all, or that when apparent it is non-uniform in distribution. This seems somewhat consistent with the fact that self-efficacy and interest relationships to achievement are not hugely powerful. Thus, it seems plausible that the detected DIF is more a case of real differences rather than bias (Zumbo, 2007). This indeed may be a case in which DIF arises from a construct-relevant source.

The pattern of DIF was not identical for the SCoA items between the two subjects tested (i.e., reading and mathematics), suggesting a possible influence of the subject domain on how students with different levels of interest or self-efficacy respond to the SCoA. In terms of uniform DIF, one might conclude from MIMIC-interaction that there is no difference by subject. Whereas, the MIMIC-PA approach suggested subject

differences in terms of establishing an anchor item and the number of items with large DIF. In contrast, non-uniform DIF was seen in three items in mathematics compared to one item in reading. What it is about these subjects that causes different patterns is not clear. This discrepancy may be a function of pedagogical practices in the two subjects. It seems plausible highly positive students could be aware of the greater use of within-class group work and focus on enjoying literature/reading within English classes; whereas, mathematics classes may seem, to students with lower self-efficacy and interest, to focus on accurate performance on tests and pedagogical calls to study and correct mistakes on tested performance. However, these are highly speculative interpretations, based on stereotypes of subject classroom practices, and are not supported with information about actual practices. Nonetheless, these discrepancies raise questions about the subject-based practices around assessment.

The study extends the body of work around the MIMIC-PA and MIMIC-interaction approaches to DIF using real, rather than simulated, data. This study shows that, arguably there is no uniform DIF within the SCoA inventory according to student interest or self-efficacy. It seems, combined with published studies that have indicated metric and scalar equivalence for the SCoA (Brown, Peterson, & Irving, 2009), that the amount and strength of DIF for SCoA items do not prevent it from being used as a research instrument, at least among New Zealand secondary school students. The items with non-uniform DIF according to MIMIC-interaction analysis appear to function in accordance with how interest and self-efficacy relate to self-regulation of learning and achievement. It seems likely that the DIF maximizes the effect of these control and competence variables and, thus, legitimates the SCoA as a measure of important learning related beliefs. Further, the study shows that the MIMIC-interaction approach detects fewer DIF items than the MIMIC-PA method, despite their shared origins. Whether such difference is due to greater control over Type I error, is beyond the scope of this study and should be investigated in future studies. This paper provides further evidence for the continued use of the SCoA as a research tool within educational psychology.

References

Alexander, P. A. (1995). Superimposing a situation-specific and domain-specific perspective on an account of.. *Educational Psychologist, 30*(4), 189-193. doi:10.1207/s15326985ep3004_3

Alexander, P. A. (2003). The development of expertise: The journey from acclimation to proficiency. *Educational Researcher, 32*(8), 10-14.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice-Hall.

Bong, M. (2013). Self-efficacy. In J. Hattie & E. M. Anderman (Eds.), *International guide to student achievement* (pp. 64-66). New York: Routledge.

Brown, G. T. L. (2008). Students' conceptions of assessment inventory (SCoA Version VI) [Measurement instrument]. Auckland, NZ: University of Auckland. doi:10.17608/k6.auckland.4596820.v1

Brown, G. T. L. (2011). Self-regulation of assessment beliefs and attitudes: a review of the Students' Conceptions of Assessment inventory. *Educational Psychology, 31*(6), 731-748. doi:10.1080/01443410.2011.599836

Brown, G. T. L., & Hirschfeld, G. H. F. (2008). Students' conceptions of assessment: Links to outcomes. *Assessment in Education: Principles, Policy & Practice, 15*(1), 3-17. 10.1080/09695940701876003

Brown, G. T. L., & Walton, K. F. (2017). The effect of conceptions of assessment upon reading achievement: An evaluation of the influence of self-efficacy and interest. *Interdisciplinary Education and Psychology, 1*(103) doi:10.31532/InterdiscipEducPsychol.1.1.003

Brown, G. T. L., Irving, S. E., Peterson, E. R., & Hirschfeld, G. H. F. (2009). Use of interactive-informal assessment practices: New Zealand secondary students' conceptions of assessment. *Learning and Instruction, 19*(2), 97-111. doi:10.1016/j.learninstruc.2008.02.003

Brown, G. T. L., Peterson, E. R., & Irving, S. E. (2009). Beliefs that make a difference: Adaptive and maladaptive self-regulation in students' conceptions of assessment. In D. M. McInerney, G. T. L. Brown & G. A. D. Liem (Eds.), *Student perspectives on assessment: What students can tell us about assessment for learning.* (pp. 159-186). Charlotte, NC US: Information Age Publishing.

Brown, G. T. L., Peterson, E. R., & Yao, E. S. (2016). Student conceptions of feedback: Impact on self-regulation, self-efficacy, and academic achievement. *British Journal of Educational Psychology, 86*(4), 606-629. 10.1111/bjep.12126

Bulut, O., & Suh, Y. (2017). Detecting DIF in multidimensional assessments with the MIMIC model, the IRT likelihood ratio test, and logistic regression. *Frontiers in Education, 2*(51), 1-14. doi: 10.3389/feduc.2017.00051

Choi, S. W., Gibbons, L. E., & Crane, P. K. (2011). Lordif: A R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulation. Journal of Statistical Software, 39(8), 1-30. doi:10.18637/jss.v039.i08

DeMars, C., E. (2015). Modeling DIF for simulations: Continuous or categorical secondary trait? *Psychological Test and Assessment Modeling, 57*(3), 279–300.

Gierl, M. J., Bisanz, J., Bisanz, G. L., Boughton, K. A., & Khaliq, S. N. (2001). Illustrating the utility of differential bundle functioning analyses to identify and interpret group differences on achievement testing. *Educational Measurement: Issues and Practice, 20*, 26-36. doi: 10.1111/j.1745-3992.2001.tb00060.x

Grimm, K. J., & Widaman, K. F. (2012). Construct validity. In H. Cooper (Ed.), *APA Handbook of research methods in psychology* (Vol. 1. Foundations, Planning, Measures, and Psychometrics, pp. 621-642). Washington, DC: American Psychological Association.

Hagquist, C., & Andrich, D. (2015). Determinants of artificial DIF – a study based on simulated polytomous data. *Psychological Test and Assessment Modeling, 57* (3), 342–376.

Hamilton, L. S. (1999). Detecting gender-based differential item functioning on a constructed-response science test. *Applied Measurement in Education, 12*, 211-235. doi:10.1207/S15324818AME1203_1

Hattie, J. A. C., Brown, G. T. L., Keegan, P. J., MacKay, A. J., Irving, S. E., Cutforth, S., et al. (2005). *Assessment Tools for Teaching and Learning asTTle) Manual (Version 4, 2005).* Wellington, NZ: University of Auckland/ Ministry of Education/ Learning Media.

Hirschfeld, G. H. F., & Brown, G. T. L. (2009). Students' conceptions of assessment: Factorial and structural invariance of the SCoA across sex, age, and ethnicity. *European Journal of Psychological Assessment, 25*(1), 30-38. doi:10.1027/1015-5759.25.1.30

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp. 129-145). Hillsdale, NJ: Erlbaum Associates.

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a Model with Multiple Indicators and Multiple Causes of a Single Latent Variable. *Journal of the American Statistical Association, 70*(351), 631-639. 10.2307/2285946

Klieme, E., & Baumert, J. (2001). Identifying national cultures of mathematics education: Analysis of cognitive demands and differential item functioning in TIMSS. *European Journal of Psychology of Education, 16*(3), 385-402. Retrieved from http://www.jstor.org/stable/23420340

Lam, T. C. M., & Klockars, A. J. (1982). Anchor point effects on the equivalence of questionnaire items. *Journal of Educational Measurement, 19*(4), 317-322.

Lee, S., Bulut, O., & Suh, Y. (2017). Multidimensional extension of multiple indicators multiple causes models to detect DIF. *Educational & Psychological Measurement, 77*(4), 545-569. doi:10.1177/0013164416651116

McMillan, J. H. (2016). Section discussion: Student perceptions of assessment. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of Human and Social Conditions in Assessment* (pp. 221-243). New York: Routledge.

Muthén, B. J. P. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *49*(1), 115-132. 10.1007/bf02294210

Muthén, L. K., & Muthén, B. O. (1998-2015). *Mplus User's Guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Otunuku, M., & Brown, G. T. L. (2007). Tongan students' attitudes towards their subjects in New Zealand relative to their academic achievement. *Asia Pacific Education Review, 8*(1), 117-128. doi:10.1007/BF03025838

Pajares, M. F. (1996). Self-efficacy beliefs in academic settings. *Review of Educational Research, 66*(4), 543-578.

Peterson, E. R., Brown, G. T. L., & Jun, M. C. (2015). Achievement emotions in higher education: A diary study exploring emotions across an assessment event. *Contemporary Educational Psychology, 42*, 82-96. 10.1016/j.cedpsych.2015.05.002

Schunk, D. H., & Zimmerman, B. J. (2006). Competence and control beliefs: Distinguishing the means and ends. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of Educational Psychology* (2nd ed., pp. 349-367). Mahwah, NJ: LEA.

Shih, C.-L., & Wang, W.-C. (2009). Differential item functioning detection using the multiple indicators, multiple causes method with a pure short anchor. *Applied Psychological Measurement, 33*, 184-199. doi:10.1177/0146621608321758

Struyven, K., & Devesa, J. (2016). Students' perceptions of novel forms of assessment In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of Human and Social Conditions in Assessment* (pp. 129-144). New York: Routledge.

Teresi, J. A. (2006). Different approaches to differential item functioning in health applications. Advantages, disadvantages and some neglected topics. *Medical Care, 44*, S152-170.

Teresi, J. A., Ramirez, M., Lai, J., & Silver, S. (2008). Occurrences and sources of differential item functioning (DIF) in patient-reported outcomes measures: Description of DIF methods and review of measures of depression, quality of life and general health. *Psychology Science Quarterly, 50*, 538-612.

Tran, T. V. (2009). *Developing cross-cultural measurement*. Oxford, UK: Oxford University Press.

Vogl, E., & Pekrun, R. (2016). Emotions that matter to achievement: Student feelings about assessment. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of Human and Social Conditions in Assessment* (pp. 111-128). New York: Routledge.

Wang, W.-C., & Shih, C.-L. (2010). MIMIC methods for assessing differential item functioning in polytomous items. *Applied Psychological Measurement, 34*, 166-180. doi:10.1177/0146621609355279

Wise, S. L., & Cotten, M. R. (2009). Test-taking effort and score validity: The influence of student conceptions of assessment. In D. M. McInerney, G. T. L. Brown & G. A. D. Liem (Eds.), *Student perspectives on assessment: What students can tell us about assessment for learning* (pp. 187-205). Charlotte, NC: Information Age Publishing.

Woods, C. M. (2009). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research, 44*, 1-27. doi:10.1080/00273170802620121

Woods, C. M., and Grimm, K. J. (2011). Testing for nonuniform differential item functioning with multiple indicator multiple cause models. *Applied Psychological Measurement,* 35, 339–361. doi:10.1177/0146621611405984

Zimmerman, B. J., & Schunk, D. H. (2004). Self-regulating intellectual processes and outcomes: A social cognitive perspective. In D. Y. Dai & R. J. Sternberg (Eds.), *Motivation, emotion and cognition*. Mahwah, NJ: Lawrence Erlbaum Associates.

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. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*, 223-233. doi:10.1080/15434300701375832

Zumbo, B. D., & Witarsa, P. M. (2004). *Nonparametric IRT methodology for detecting DIF in moderate-to-small scale measurement: Operating characteristics and a comparison with the Mantel Haenszel.* Paper presented at the American Educational Research Associate Meeting, San Diego, CA.

The first author acknowledges the insights and advice of Prof Xitao Fan, University of Macao, and Prof Pat Alexander, University of Maryland, in shaping this manuscript. Professors Wang Wen Chung (Hong Kong Institute of Education) and Shih Ching-Lin (National Sun Yat-sen University, Taiwan) are thanked for access to the MIMIC-PA syntax.

Corresponding author:
Professor Gavin T. L. Brown, PhD
School of Learning, Development and Professional Practice
Faculty of Education
The University of Auckland
Private Bag 92019
Auckland, 1142
New Zealand
gt.brown@auckland.ac.nz