Checking Equity: Why Differential Item Functioning Analysis Should Be a Routine Part of Developing Conceptual Assessments

Patrícia Martinková,[†]* Adéla Drabinová,[†‡] Yuan-Ling Liaw,[§] Elizabeth A. Sanders,[||] Jenny L. McFarland,[¶] and Rebecca M. Price[#]

[†]Institute of Computer Science, Czech Academy of Sciences, Praha 182 07, Czech Republic; [‡]Department of Probability and Mathematical Statistics, Faculty of Mathematics and Physics, Charles University, Praha 186 75, Czech Republic; [§]Center for Educational Measurement, University of Oslo, Oslo 0318, Norway; [||]College of Education, University of Washington, Seattle, WA 98195; [¶]Biology Department, Edmonds Community College, Lynnwood, WA 98036; [#]School of Interdisciplinary Arts and Sciences, University of Washington, Bothell, Bothell, WA 98011

ABSTRACT

We provide a tutorial on differential item functioning (DIF) analysis, an analytic method useful for identifying potentially biased items in assessments. After explaining a number of methodological approaches, we test for gender bias in two scenarios that demonstrate why DIF analysis is crucial for developing assessments, particularly because simply comparing two groups' total scores can lead to incorrect conclusions about test fairness. First, a significant difference between groups on total scores can exist even when items are not biased, as we illustrate with data collected during the validation of the Homeostasis Concept Inventory. Second, item bias can exist even when the two groups have exactly the same distribution of total scores, as we illustrate with a simulated data set. We also present a brief overview of how DIF analysis has been used in the biology education literature to illustrate the way DIF items need to be reevaluated by content experts to determine whether they should be revised or removed from the assessment. Finally, we conclude by arguing that DIF analysis should be used routinely to evaluate items in developing conceptual assessments. These steps will ensure more equitable—and therefore more valid—scores from conceptual assessments.

INTRODUCTION

Knowledge assessments that are used to measure students' understanding of disciplinary concepts need to produce valid and reliable scores (Downing and Haladyna, 2006; American Educational Research Association, American Psychological Association, National Council on Measurement in Education [AERA, APA, NCME], 2014). This robustness is essential for high-stakes tests used, for example, in college admissions, and it is also essential for drawing inferences about student performance on low-stakes assessments, such as those within science classrooms. For example, in biology, several concept inventories have been developed that measure what students understand about specific core concepts. A variety of methods are used to explore the validity of scores during the development of assessment tools (e.g., Libarkin, 2008; Adams and Wieman, 2011; Reeves and Marbach-Ad, 2016).

During the process of validation, developers often test for differences in performance among two or more groups of students as one way of gathering evidence of the presence or absence of test bias, such as whether men and women perform differently, whether native speakers of the testing language perform consistently better than others, or whether race/ethnicity is linked with performance (Walker and Beretvas, 2001;

Ross Nehm, *Monitoring Editor*

Submitted October 24, 2016; Revised February 21, 2017; Accepted March 15, 2017

CBE Life Sci Educ June 1, 2017 16:rm2

DOI:10.1187/cbe.16-10-0307

*Address correspondence to: Patrícia Martinková (martinkova@cs.cas.cz).

Walker, 2011). In addition, attending to the performance of different groups is critical for equity—assessments should not discriminate against any individual (e.g., Libarkin, 2008); developers must show that performance on the assessment is not related to factors that are irrelevant to the construct being tested.

In this paper, we explain how applied data analysis techniques can be used to identify potentially biased, or unfair, test items. We define differential item functioning, or DIF, and explain the suite of statistical approaches, known as DIF analysis, used to identify DIF items—those test items for which different groups of students perform differently. After explaining the theory behind DIF analyses and describing some of the most common methods, we present two scenarios with real and simulated data to demonstrate how to conduct several types of DIF analysis. The first scenario consists of data from the Homeostasis Concept Inventory (HCI; McFarland *et al.*, 2017), which was administered to students from a range of institutions throughout the United States. The second scenario consists of a simulated data set designed to show that the distribution of the total scores of two groups can be exactly the same, even when some test items are biased against one of those groups. Together, these cases show how DIF analyses provide a rich, nuanced understanding of how different groups perform on a test.

While some biology education articles have previously employed DIF and/or highlighted its importance (e.g., Penfield and Lee, 2010; Federer *et al.*, 2016; Romine *et al.*, 2016), many of the validation studies of concept inventories have not checked for DIF or potential unfairness of items. For example, none of the 22 articles in a list of biology concept inventories (SABER, n.d.) used DIF analysis to check for potential bias in their items. That said, interest in DIF analysis is growing within the biology education community. One recent paper by Deane *et al.* (2016) considered item bias, and others have been using and advocating for item-level analyses (Neumann *et al.*, 2011; McFarland *et al.*, 2017). We, along with others (AERA, APA, NCME, 2014; see also Reeves and Marbach-Ad, 2016), argue that items flagged as DIF have a strong potential to threaten the validity of scores if they are not further investigated, and therefore DIF analysis should be performed routinely when developing conceptual assessments. We conclude this paper by reviewing recent examples from biology education that used DIF analysis.

BACKGROUND

Identifying Achievement Gaps

In the high-stakes testing world, assessment researchers consistently evaluate the fairness of tests and explore the reasons behind achievement gaps (e.g., Sabatini *et al.*, 2015; Huang *et al.*, 2016). This type of detailed analysis occurs less frequently in the low-stakes testing world, and, if present, is typically restricted to comparing groups on test *total scores* (Steif and Dantzler, 2005). Differences between groups on total scores may be identified graphically with box plots, histograms, or density plots, and can also be tested empirically with *t* tests, chi-square tests, or regression models that account for miscellaneous fixed and random factors (Gelman and Hill, 2007; Moore *et al.*, 2015).

In contrast to examining group differences on total scores, examining differences at the item level provides clarity as to where exactly group differences are located and whether there is any pattern in those differences. The simplest item-level analysis involves comparing the groups on the proportion of examinees who answer the item correctly (this proportion is called "difficulty" and denoted as p in psychometric classic test theory literature; Allen and Yen, 1979). However, group differences in item difficulty can be due to real and important group differences or may be blurred by the fact that one group has higher knowledge of the tested topic overall. Therefore, additional analyses are required that consider both the total score and individual item performance simultaneously.

DIF analysis encompasses a set of approaches for comparing performance of groups on individual items while simultaneously considering the students' potential to score well on the test (Holland and Wainer, 1993; Camilli, 2006; Zumbo, 2007; Magis *et al.*, 2010). Therefore, DIF analysis is more useful than comparing total scores for identifying potential unfairness and for assessing the causes of achievement gaps (Zieky, 2003). As we will demonstrate, DIF analysis can identify achievement gaps that are not revealed when comparing total scores (see *Case Studies*). To our knowledge, DIF analysis has rarely been used in the development of low-stakes tests such as concept inventories (but see McFarland *et al.*, 2017, for an example of DIF analysis).

DIF analysis is conducted by comparing a reference group (typically the majority, or normative, group) with a focal group (typically the minority, or disadvantaged, group). For example, a study probing for bias against women would use men as the reference group and women as the focal group. Similarly, when testing whether items are biased against a historically underrepresented minority group on an assessment developed in the United States, we would consider white students as the reference group, and we might consider African-American students to be the focal group. In a similar vein, native language speakers being assessed would be considered the reference group, with language learners as the focal group.

An Item Can Measure More Than What Was Intended

DIF occurs when an item measures more than one underlying latent trait and when cognitive differences exist on one of these other, so-called secondary, latent traits (Ackerman, 1992; Shealy and Stout, 1993; Roussos and Stout, 1996). A latent trait (also known as latent knowledge, latent ability, or, more generally, latent variable) is an individual's true knowledge or understanding of the construct being measured, and it can be estimated but not directly measured. The simplest estimate of the latent trait is total score. When developing biology concept inventories, for example, the goal is that the items only measure students' biological concept knowledge (primary latent trait), and that additional cultural knowledge (secondary latent trait related to identity) is not necessary to answer items correctly. In other words, we wish to know whether the items only measure students' knowledge of information relevant to the concept. The presence of DIF for a given item would indicate that the item may measure a secondary latent trait, either alone (completely missing the target concept) or in concert with the primary trait (which requires knowledge of the target concept and the secondary concept). For example, suppose a student needed both an understanding of the concept of homeostasis and knowledge of difficult English vocabulary (e.g., to understand the phrase "hypertension is characteristic of diabetic nephropathy"), that is not essential for understanding the concept of homeostasis. Moreover, if the focal group of students does not have the requisite knowledge of English, then the focal group would be more likely to answer incorrectly compared with those in the reference group *even if the focal group has the same level of homeostasis knowledge as the reference group*. Moreover, a test containing items exhibiting DIF could in turn create inaccurate observed total scores, resulting in inaccurate estimation of the focal group's primary latent trait (e.g., biological concepts).

Fair or Unfair?

Although the presence of DIF is a signal that an item may be biased, it does not guarantee that the item is unfair. Rather, the presence of DIF indicates the existence of a latent trait besides the one of primary interest. Fairness is established subsequently if the secondary latent trait that was detected statistically is intentionally related to the primary latent trait. It is possible that the secondary latent trait is required by the content and the test specifications, even if the reference and focal groups perform differently.

An example of a situation in which a primary and secondary latent trait are both required for a test occurred recently in a biology admission test for medical school in the Czech Republic (Drabinová and Martinková, 2016; see also Štuka *et al.*, 2012). On one item about childhood illnesses, DIF analysis revealed that women performed better than men. Content experts reviewed the item, and concluded that the difference occurred because women in the Czech Republic spend more time with children than men and therefore have more experience with childhood illnesses (Drabinová and Martinková, 2016). The faculty, however, still considered the item to be fair despite the gender difference, because medical experts need to be familiar with childhood illnesses. In this case, the test writers decided that the secondary latent trait, knowledge of childhood illnesses, was related to the primary concept being tested, which was biology in medicine.

Other clear examples of fair items that exhibited DIF exist in the literature (Doolittle, 1985; Hamilton, 1999; Zenisky *et al.*, 2004; Liu and Wilson, 2009; Kendhammer *et al.*, 2013). Even if the item flagged as DIF is later reviewed and considered fair, the act of identifying these gaps in conceptual understanding can inform teaching and, subsequently, help educators and policy makers to reduce such gaps in the future.

The Czech biology medical test shows how critical it is for content experts to review whether DIF is the result of unintended secondary latent traits (see also Ercikan *et al.*, 2010). Only if the presence of DIF can be attributed to unintended item content (e.g., related to cultural background) or some other unintended item property (e.g., method of test administration) is the item said to be unfair. In such cases, content that is unrelated to the concept being tested increases the likelihood an individual will answer the item correctly.

Items that have been evaluated with multiple rounds of DIF analysis and content expert adjustments can help to decrease unfair achievement gaps (e.g., Penfield and Lee, 2010). For example, Siegel (2007) demonstrated ways to clarify item wording for second language learners so that they were fairly tested on the content rather than their language. In other words, ensuring that the items are clearly worded bolsters our confidence that we are truly assessing what students know. Martinello and Wolf (2012) demonstrated three situations in which individuals from focal groups responded incorrectly to math items that they were able to answer correctly during interviews. In one of the examples, a high school student from another country who was still learning English (the language of the test) did not understand words with multiple meanings the same way that native speakers would, particularly when the words used in the problem were culture bound. For example, the word "tip" can refer to tipping a waiter, but in other circumstances it means the top of an object, not a percentage of money. Thus, questions that ask students to calculate a tip for a waiter can be unfair (Martinello and Wolf, 2012). Some students will answer incorrectly even if they can demonstrate the knowledge required to do the task, in this case approximating the percentage of another number. Unfair items translate into unfair reflections of an individuals' true ability or knowledge, and they also have the strong potential to discourage students from underrepresented groups from becoming interested in a subject (Wright *et al.*, 2016). Thus, using DIF analysis to identify DIF items that are unfair enables us to reformulate or remove them.

METHODS FOR DETECTING DIF

In this section, we review the most commonly used statistical methods that have been developed to detect DIF (Holland and Wainer, 1993; Millsap and Everson, 1993; Camilli and Shepard, 1994; Clauser and Mazor, 1998; Magis *et al.*, 2010). We focus on methods for tests with dichotomous items, which include binary items graded as true (1) or false (0), or as correct (1) or incorrect (0), on multiple-choice or free-recall tests. Methods for detecting DIF on other types of items (e.g., those graded on a rating, ranking, or partial-credit scale) are similar but beyond the scope of this paper. Generally speaking, statistically detecting items exhibiting DIF requires that we match students on relevant knowledge (e.g., using their total scores on the assessment being evaluated as an estimate of ability, or latent trait), and then test whether students who are matched for ability but from different groups perform similarly on a given item.

The methods for detecting DIF vary depending on how students are matched. Classical methods (e.g., Mantel-Haenszel statistic and logistic regression) match students based on their total scores; methods based on item response theory (IRT) models, such as the Wald χ^2 test (also known as Lord's test; Lord, 1980) and Raju's area test, consider student ability as a latent variable estimated together with item parameters in the model (Hills, 1989; Millsap and Everson, 1993; Camilli and Shepard, 1994). Generally, IRT methods are computationally more demanding and require larger sample sizes. However, IRT methods are more precise than others, because they more accurately estimate the latent trait instead of using total score as the proxy.

Here, we describe the three most common approaches for detecting DIF (Table 1). First, we discuss the Mantel-Haenszel χ^2 test, which functions well even for very small sample sizes (Mantel and Haenszel, 1959) and allows researchers to calculate statistics quickly using basic arithmetic. Second, we describe procedures that rely on logistic regression (Zumbo, 1999), which provides a more precise description of DIF compared with the Mantel-Haenszel procedure, and allows for

TABLE 1. An overview of the methods for detecting DIF presented in this study

Method	Strengths	Limitations
Classical methods		
Mantel-Haenszel χ^2 test	Easy to calculate by hand Can handle small sample sizes	Not always able to detect nonuniform DIF
Logistic regression		
Two-parameter	Able to capture nonuniform DIF	Does not account for guessing
Three-parameter	Accounts for guessing	Some convergence issues can be observed
IRT-based methods		
Three-parameter logistic (3PL) IRT Wald test	More precisely estimates latent ability	Requires large sample size ($N > 500$ in each group)

distinguishing between two types of DIF: uniform and nonuniform DIF. Finally, we discuss methods based on IRT models (Lord, 1980; Raju, 1990; Thissen *et al.*, 1994), which more accurately estimate both item characteristics and student abilities but require relatively larger sample sizes.

Mantel-Haenszel χ^2 Test: Method for Small Samples

The Mantel-Haenszel test is an extension of the test χ^2 for contingency tables (Agresti, 2002) that sorts students into groups based on their total scores, k (Mantel and Haenszel, 1959; Holland, 1985; Holland and Thayer, 1988). For a given item and a given level of total score k, a 2 × 2 contingency table is created (Table 2).

Table 2 enumerates the number of students with total score k from the reference group who answered the item correctly (A_k) and incorrectly (B_k), and the respective number of students from the focal group (C_k, D_k). The item is not DIF if the odds of answering the item correctly (at a given total score level k) are about the same across focal and reference group, that is, if the odds ratio, $\alpha_k = \frac{A_k/B_k}{C_k/D_k} = \frac{A_k D_k}{B_k C_k}$, is close to 1.

To account for all total score levels simultaneously, the Mantel-Haenszel estimate of the odds ratio α_{MH}, can be calculated as the weighted average of the odds ratios across the score levels:

$$\alpha_{\text{MH}} = \frac{\sum_k \frac{A_k D_k}{N_k}}{\sum_k \frac{B_k C_k}{N_k}}$$

If the item functions the same for different groups for all levels of total score, k, then the odds ratio α_{MH} equals 1, indicating that there is no DIF. For a DIF item that favors the reference group, α_{MH} is greater than 1. When a DIF item favors the focal group, α_{MH} is less than 1.

TABLE 2. Contingency table for one item and level of total score equal to *k*

	Correct answer	Incorrect answer	Total
Reference group	A_k	B_k	$A_k + B_k$
Focal group	C_k	D_k	$C_k + D_k$
Total	$A_k + C_k$	$B_k + D_k$	$N_k = A_k + B_k + C_k + D_k$

To test for the presence of DIF, the Mantel-Haenszel χ^2_{MH} statistic can be calculated as

$$\chi^2_{\text{MH}} = \frac{\left\{\left|\sum_k \left[A_k - \frac{(A_k+B_k)(A_k+C_k)}{N_k}\right]\right| - 0.5\right\}^2}{\sum_k \frac{(A_k+B_k)(A_k+C_k)(B_k+D_k)(C_k+D_k)}{N_k^2(N_k-1)}}$$

and compared with a critical value from the χ^2 distribution with $df = 1$ to determine the p value (Mantel and Haenszel, 1959; Holland and Thayer, 1988). As an alternative, the Mantel-Haenszel estimate of the odds ratio α_{MH} is sometimes converted to a standard metric called the delta scale (Zieky, 1993):

$$\Delta_{\text{MH}} = -2.35\ln(\alpha_{\text{MH}}).$$

In addition, Δ_{MH} can be assigned to one of three categories: A—negligible difference, B—moderate difference, or C—large difference. At one extreme, category A contains items with Δ_{MH} not significantly different from zero (using a test based on the normal distribution; see Agresti, 2002), or with small effect size, $|\Delta_{\text{MH}}| < 1$. At the other extreme, category C contains the items with $|\Delta_{\text{MH}}|$ significantly greater than one and large effect size, $|\Delta_{\text{MH}}| \geq 1.5$. Category B consists of all other items.

Significance testing of χ^2_{MH} or Δ_{MH} results in making multiple comparisons, because every item on the instrument is tested. Therefore, to avoid inflating the type I error rate beyond the nominal level (i.e., minimizing false discoveries of DIF), Kim and Oshima (2013) recommended using the Benjamini-Hochberg p value correction, a sequential multiple comparison procedure employing the Dunn-Šidák adjusted critical p value computation (Benjamini and Hochberg, 1995). This procedure maximizes power while controlling the false discovery rate to the nominal value (typically 5%).

Methods Using Logistic Regression: Looking Closer at Item Functioning

Another method for detecting DIF for binary-scored items (i.e., 1 = correct and 0 = not correct) proposed by Swaminathan and Rogers (1990) uses logistic regression to model each item individually (see also Zumbo, 1999; Agresti, 2002; Magis *et al.*, 2010). This method predicts the probability that student i answers item j correctly (i.e., $Y_{ij} = 1$), conditional on the total score X_i as follows:

$$P\left(Y_{ij} = 1 | X_i\right) = \frac{1}{1 + e^{-(\beta_{0j} + \beta_{1j} * X_i)}} \qquad \text{(Model 1)}$$

The two parameters β_{0j}, β_{1j}, describe properties of item *j*, and they are estimated from the model. Parameter β_{0j} is an intercept, that is, the probability of answering the item correctly for students with a total score of zero (note that if the total score is centered or standardized around zero, then the intercept would indicate the probability of answering the item correctly for students with an average total score). Parameter β_{1j} represents the effect of the total score on the intercept (again, the interpretation of this effect depends on how the total score is scaled), that is, β_{1j} is the effect on the probability of answering each particular item correctly for each one-unit increase in the total score. Note that the model parameters are typically estimated by taking the natural log of the odds of the probability of answering the item correctly, and thus the estimated parameter values given in most software outputs are typically in "logits," resulting in the model name, "logistic" regression. The estimated logistic regression line relating the total score to the probability of answering the item correctly is often called the item characteristic curve.

To test for DIF, the linear term $\beta_{0j} + \beta_{1j} * X_i$ in the logistic regression model 1 needs to be extended by allowing the parameters to differ by group, G_i, as follows:

$$P\left(Y_{ij} = 1|X_i, G_i\right) = \frac{1}{1+e^{-(\beta_{0j}+\beta_{1j}*X_i+\beta_{0DIFj}*G_i+\beta_{1DIFj}*X_i*G_i)}} \qquad \text{(Model 2)}$$

The new parameters $\beta_{0DIF\,j}$, β_{1DIFj} describe the potential differences in intercept and slope values between the focal and reference groups ($\boldsymbol{G}_i$). If neither of these parameter estimates is statistically significant, it is concluded that there is no DIF present in the item (Figure 1A). Estimates of the slopes and intercepts in a logistic regression model are often estimated using iterative weighted least squares (Agresti, 2002). The significance of the parameters involving differences between groups (and thus the detection of DIF) is performed either by comparing the more complex model (which allows groups to differ on the intercept, Figure 1B, or intercept and slope, Figure 1C) with the simpler model (which constrains groups to have the same intercept and slope, Figure 1A) using a likelihood ratio test or by conducting a Wald test on each estimate (see Agresti, 2002). In either case, the Benjamini-Hochberg correction for multiple comparisons would need to be applied to determine the critical *p* value for detecting DIF.

Uniform and Nonuniform DIF. The logistic regression model allows us to distinguish between two types of DIF: uniform DIF, which affects students at all levels of the total score in the same way (Figure 1B), and nonuniform DIF, which affects students in specific ranges of the total score inconsistently (Figure 1C). If groups only differ significantly on the intercept, that is, if $\beta_{0DIF\,j}$ is nonzero, then the item is said to have uniform DIF. With uniform DIF, the item characteristic curves for the two groups have the same shape and do not cross (Figure 1B). Such items favor one group over another group across the entire range of the total score, albeit less so at the extreme ends of the total score distribution.

However, if groups differ on their slopes (i.e., if $\beta_{1DIF\,j}$ is nonzero), the item is said to have nonuniform DIF. When items have nonuniform DIF, the item characteristic curves for different groups have different shapes, and these curves cross (Figure 1C). Such items favor one group within a specific range of the total score, and then, at some point along the total score distribution, the difference flips to favor the other group.

Shifting from Logistic Regression toward IRT Models

The Two-Parameter Model Reparameterized. Before explaining IRT models, it is helpful to describe how the basic two-parameter logistic regression model for an item (model 1) can also be fitted using different set of parameters:

$$P\left(Y_{ij} = 1|Z_i\right) = \frac{1}{1+e^{-a_j(Z_i-b_j)}} \qquad \text{(Model 3)}$$

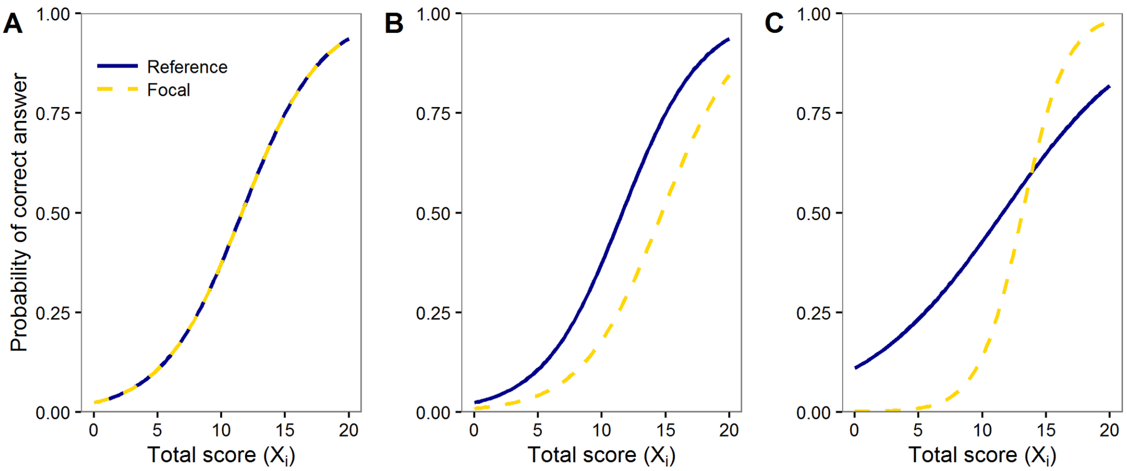

FIGURE 1. Characteristic curves for reference (blue solid) and focal (yellow dashed) group. (A) The shape and placement of the curves are identical, so there is no DIF. (B) The item shows uniform DIF between the reference and focal group. (C) The item shows nonuniform DIF between the reference and focal group: the reference group has the advantage below the total score of 14, and the focal group has the advantage for total scores above 14.

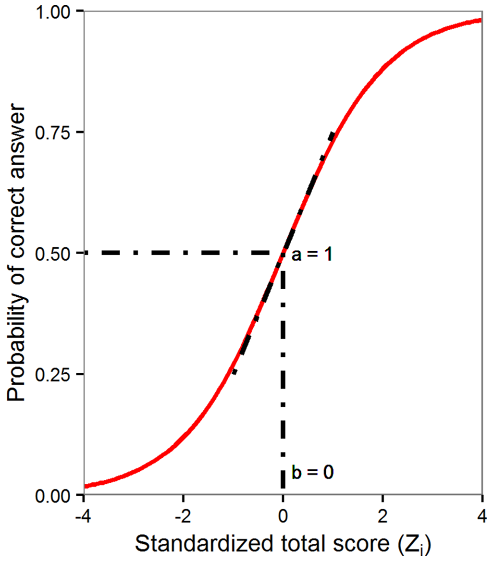

FIGURE 2. Characteristic curve of logistic regression model 3. The line representing the probability of a student answering the item correctly is plotted against the standardized total score (Z_i). Parameter *b* represents difficulty (location of inflection point); parameter *a* is discrimination (slope).

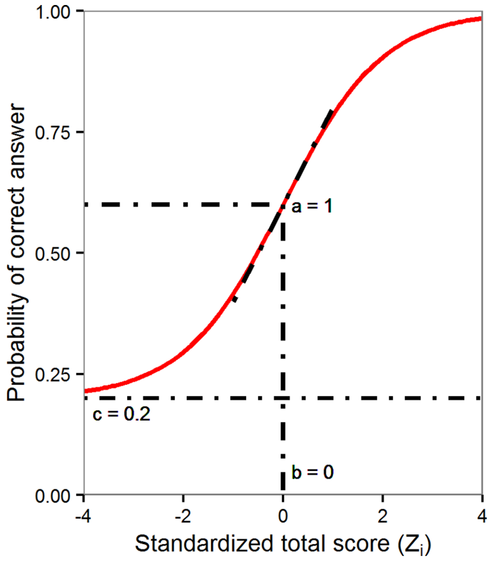

FIGURE 3. Characteristic curve of three-parameter logistic regression model 4. The guessing parameter (*c*) is the probability that item is guessed without necessary knowledge, and it is represented as the lower asymptote. The inflection point now occurs at the standardized total score, where the probability of the correct answer is (1 + c)/2.

Model 3 differs from model 1 in two ways. First, the standardized total score, Z_i, is always used instead of the total score X_i, which makes changes in all parameter estimates that had relied on the total score to become reinterpreted in terms of standard deviations from the mean total score (note that the total score could also be scaled in standard deviations in model 1). Second, and more crucially, instead of the parameters β_{0j}, β_{1j} used in logistic regression, two different parameters, a_j, b_j, are estimated. The parameters b_j and a_j describe the difficulty and discrimination of item *j*, respectively. In this new model, the terms "difficulty" and "discrimination" are used slightly differently than they are usually used in classic measurement theory research (Allen and Yen, 1979), although conceptually they are similar. The difficulty term, b_j, is the standardized total score that is needed to answer item *j* correctly with 50% probability. In addition, b_j is the inflection point on the item characteristic curve (Figure 2). The discrimination term, a_j, is the slope of the curve at the inflection point. We follow the same procedures to test the significance of these terms as was described for the other logistic regression models, using a likelihood ratio test or a Wald test.

The Three-Parameter Model: Accounting for Guessing. The two-parameter logistic regression models discussed so far (models 1 and 3) did not account for the possibility that students may correctly answer an item simply by guessing, which is an expected behavior, especially for multiple-choice tests. To account for guessing, we can extend the two-parameter model (model 3) to include a guessing parameter, c_j, as follows:

$$P\left(Y_{ij}=1|Z_i\right)=c_j+\left(1-c_j\right)\frac{1}{1+e^{-a_j\left(Z_i-b_j\right)}} \qquad \text{(Model 4)}$$

Technically speaking, the guessing parameter, c_j, actually captures pseudo-guessing, which takes into account the probability of choosing each of the alternative response options (also known as distractors) rather than assuming an equal probability across all option choices. As before, the parameters b_j and a_j again describe the difficulty and discrimination of *j*th item, respectively. In the item characteristic curve, the new pseudo-guessing parameter, c_j, is represented by the lower asymptote (Figure 3). Note that, for the case where c_j is assumed to be 0, the model reduces to the previous two-parameter logistic regression, model 3.

As with our original two-parameter logistic regression that tested for DIF (model 2), we can test the effect of group membership, G_i, on item parameters to determine the presence of DIF by adding the parameters a_{DIFj} and b_{DIFj} to the three-parameter logistic model 4 as follows (Drabinová and Martinková, 2016):

$$P\left(Y_{ij}=1|Z_i,G_i\right)=c_j+\left(1-c_j\right)\frac{1}{1+e^{-\left(a_j+a_{\mathrm{DIF}j}*G_i\right)\left(Z_i-\left(b_j+b_{\mathrm{DIF}j}*G_i\right)\right)}}$$

(Model 5)

The new parameters b_{DIFj} and a_{DIFj} are the differences in difficulty and discrimination, respectively, between the focal and reference group, and parameter c_j accounts for the possibility of guessing on the *j*th item. These parameters are estimated using the nonlinear least-squares method, and DIF for each item is detected with an *F*-test of the submodel (i.e., the model without group membership included) or using likelihood ratio tests (Dennis *et al.*, 1981); as with the other models, the Benjamini-Hochberg correction for multiple comparisons can be applied to control for type I error inflation.

IRT Models: Assuming a Latent Trait

IRT models, sometimes called latent trait models, are similar to logistic regression models in that they also predict the probability of a student answering an item correctly as a function of both the student ability and the item parameters. However, IRT models are more precise, in that students' true knowledge (theta) is considered latent—or unobserved—because it follows some distribution (e.g., normal) and can only be estimated from performance on observed indicators, in this case items.

The three-parameter logistic IRT model 6 (also called the 3PL IRT model) is analogous to the three-parameter logistic regression model 4, with θ representing the true, unobserved construct level and replacing Z (the standardized observed total score) as

$$P\left(Y_{ij} = 1 | \theta_i\right) = c_j + \left(1 - c_j\right) \frac{1}{1 + e^{-a_j\left(\theta_i - b_j\right)}} \qquad \text{(Model 6)}$$

The interpretation of the item parameters a_j, b_j, and c_j, remains the same as in the three-parameter model 4. However, in IRT models, θ has a specific distribution that is estimated together with item parameters.

Extending IRT Models to Test for DIF. As with the logistic regression approaches we outlined in models 2 and 4, in IRT models, interactions between group membership G and item parameters b_{DIFj} and a_{DIFj} can be added to the model to test for DIF. The parameters are then estimated jointly for all items simultaneously; this is often done using marginal maximum likelihood (Magis *et al.*, 2010), although other estimation algorithms are possible. Commonly, likelihood ratio tests are used to test whether the model that allows groups to differ on item characteristics fits better than the simpler model that constrains groups to have the same item characteristics (Thissen *et al.*, 1994). In addition, the Wald χ^2 test of differences in parameters between groups (Figure 4A) or Raju's test (Figure 4B) of the differences in areas between groups' characteristic curves (Raju, 1988, 1990) can be used to evaluate the presence of DIF. We use the Wald test in the case studies presented in this paper and then apply the Benjamini-Hochberg critical p value correction for multiple comparisons.

Other IRT Models. There are other IRT models as well. First, if the guessing parameter of model 6 is fixed at $c_j = 0$, then the model reduces to what is called a two-parameter IRT model, for which logistic regression model 3 is a proxy. Second, if the discrimination parameter is fixed at $a_j = 1$, in addition to constraining the guessing parameter to 1, then model 6 reduces to a (one-parameter) Rasch model (Rasch, 1960). Each of these two simpler models can also be extended to account for DIF just as outlined above.

Rasch models have recently received a great deal of attention in biology education research, including the recommendation that they be used more frequently for assessment development (Boone, 2016). The biggest advantage of Rasch models, like other IRT models, over the classical test theory models (using only total scores) is that they allow us to estimate the relationship between student ability and all item difficulties. Moreover, this relationship can be visualized using a type of graph known as a person-item map (also called the Wright map).

Although we agree with Boone (2016) in advocating for use of person-item maps (and, in fact, have included a person-item map in our analysis of the HCI; McFarland *et al.*, 2017), we also acknowledge the limits of the one-parameter IRT model: constraining all items to have the same discrimination levels and disallowing for guessing. Hence, we describe the three-parameter IRT model in this paper to afford researchers with a more flexible model that not only provides more information about discrimination and guessing parameters but also allows us to know whether groups differ on these parameters.

Choosing a Model with Sample Size in Mind

One limitation of IRT models is that they require a relatively large sample size. For example, it is recommended that data be collected for 500 students in the reference group and 500 students in each of the focal groups for fitting and calibrating items parameters in the three-parameter IRT model (Kim and Oshima, 2013). Nevertheless, as already mentioned, the strength of two- or three-parameter IRT models is that they can allow for additional information, such as discrimination and guessing, to be estimated and used to inform assessment development

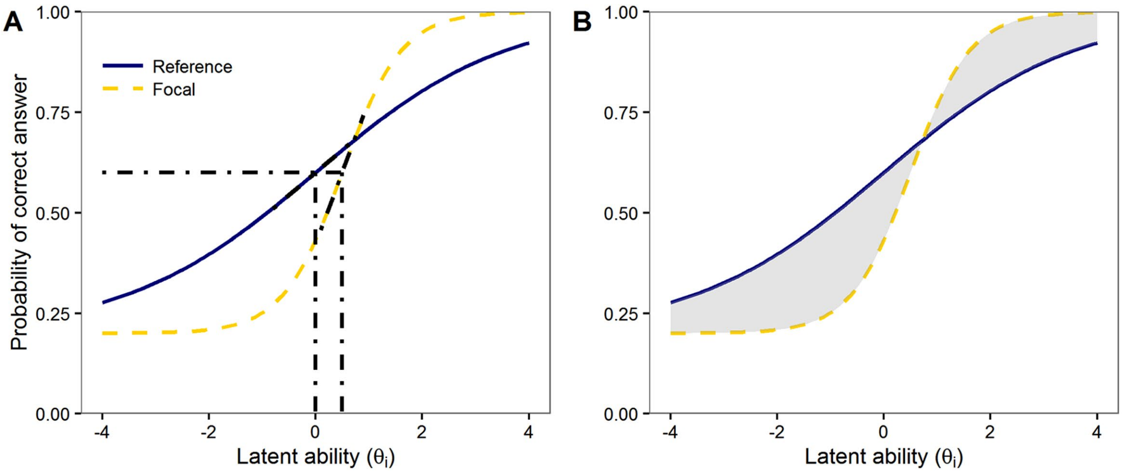

FIGURE 4. IRT methods for detecting DIF. (A) Wald χ^2 statistic is based on differences in parameter estimates for the two groups. (B) Raju's test is based on area between IRT characteristic curves for the two groups.

(Holland and Wainer, 1993; Camilli, 2006; Zumbo, 2007; Magis *et al.*, 2010).

Software Options

DIF testing using logistic regression analysis (e.g., models 2 and 5) can be carried out in a variety of widely available general statistical analysis software, such as R (R Core Team, 2016), SAS (SAS Institute, 2013), SPSS (IBM, 2013), STATA (StataCorp, 2015), and others. For DIF analysis within an IRT model, there are several commercially available packages, including Winsteps (Linacre, 2005), IRTPRO (Cai *et al.*, 2011), and ConQuest (Wu *et al.*, 1998); for other psychometric software, see www.crcpress.com/Handbook-of-Item-Response-Theory-Three-Volume-Set/Linden/p/book/9781466514393. Although each of these software packages has its own strengths, we note that the Rasch (one-parameter) IRT models are limited to testing uniform DIF due only to their simpler nature (a two-parameter model would be required to test nonuniform DIF). As such, software that estimates two-parameter IRT models is required for testing nonuniform DIF. We also note that the cost of commercially available software can sometimes be a barrier to researchers. Thus, in this paper, we illustrate examples using a freely available and flexible interactive online interface application called ShinyItemAnalysis (Martinková *et al.*, 2017), which was developed within the freely available statistical software environment, R (R Core Team, 2016) and its libraries (e.g., difR, by Magis *et al.*, 2016; difNLR by Drabinová *et al.*, 2017). The Shiny application provides a Web-based graphical user interface that makes it straightforward for users to work with R. The ShinyItemAnalysis package makes use of that interface to provide an easy to implement, user-friendly software for test and item analysis, including detection of DIF (Martinková *et al.*, 2017). We also provide R code for examples from this paper in the Supplemental Material.

CASE STUDIES

We use two data sets to provide context for and illustrate the use of DIF analysis for flagging potentially biased items. These case studies were selected to emphasize the fact that inferences about test fairness based on total scores alone may be misleading. In both case studies, the conclusions that would have been drawn solely from comparing test scores between different groups differ from the conclusions drawn from DIF analysis. Our examples purposefully illustrate two extremes to explain why analyzing total scores is not sufficient for assessing fairness. At one extreme, using a DIF analysis of the HCI (McFarland *et al.*, 2017), we observe a gap in total scores between two groups even though there is no DIF. At the other extreme, we employ a simulated data set to illustrate a case in which the distribution of total scores of two groups is exactly the same, but DIF exists. The simulated data set is a particularly powerful example because it shows that it is theoretically possible to detect DIF even when the distribution of total scores of two groups is exactly the same. This outcome is admittedly improbable, but considering this theoretical possibility helps explicate the strength of DIF analysis. This section ends with a brief overview of other studies that have used DIF analysis in biology, including a discussion of how this result leads to the reformulation of items and, therefore, an assessment that is more equitable and fair.

Case 1: HCI Data Set

The first data set was collected during the final validation of the HCI (McFarland *et al.*, 2017) and illustrates that finding a difference in total score between two groups does not necessarily indicate item bias. The HCI is a 20-item multiple-choice instrument designed to measure undergraduate student understanding of homeostasis in physiology. The HCI was validated with a sample of 669 undergraduate students, out of whom 246 identified themselves as men, 405 identified themselves as women, and the rest did not respond to this question (McFarland *et al.*, 2017). While the overall sample of 669 students is large, we knew that the sample sizes of the two subgroups might be small enough ($ns < 500$; in each group $n < 500$) that IRT models would be underpowered.

In the HCI data set, we observed a statistically significant gender gap in total scores, with men performing better (two-sample t test $p < 0.01$; Figure 5A). The average total score was 12.70 for men (SD = 3.74) and 11.92 for women (SD = 3.55). However, subsequent analysis using the Mantel-Haenszel test, the logistic regression, and the Wald test based on a three-parameter IRT model revealed no significant DIF items (see Supplemental Table 1). We therefore concluded that the HCI test is fair and that the difference between the groups on the total score was due to differences in how women and men understand the concept being targeted (i.e., a true achievement gap), rather than differences in additional content necessary to understand the items. In other words, the gender gap in the total score represents a real difference in understanding, not items that unfairly favor men over women.

This case study also provides an example of how it can be challenging to fit a three-parameter IRT model to a small sample. For men ($n = 246$), the initial model yielded unusually large standard errors for parameter estimates for item 17. This was a particularly difficult item based on a common misconception (see also McFarland *et al.*, 2017). But the large standard errors were more likely due to the fact that the sample size was relatively small for such a complex IRT model. In the end, we removed item 17 and then were able to get a good fit with the model.

Case 2: Simulated Data Set

The second data set is a simulated data set of 1000 men and 1000 women taking a 20-item, binary test inspired by Graduate Management Admission Test (GMAT; Kingston *et al.*, 1985, p. 47). This data set was designed to illustrate that DIF items may be present even when different groups have exactly the same distributions of total scores. We generated a data set in which the distribution of total scores was identical for men and women (Figure 5B), even though they performed differently on the first two items of the test. The way in which we constructed this data set (see Supplemental Tables 2 and 3) guaranteed that item 1 would have uniform DIF (Figure 6A) and item 2 would have nonuniform DIF (Figure 6B). The Mantel-Haenszel test, the logistic regression models, and the IRT models all flagged the first two items correctly as DIF (Supplemental Table 4).

Note that this example illustrates item characteristic curves that distinguish between uniform and nonuniform DIF. We specifically used the three-parameter logistic regression (model 5) to generate these curves (Figure 6 and Supplemental Table 4), because we wanted to take into account guessing, as guessing

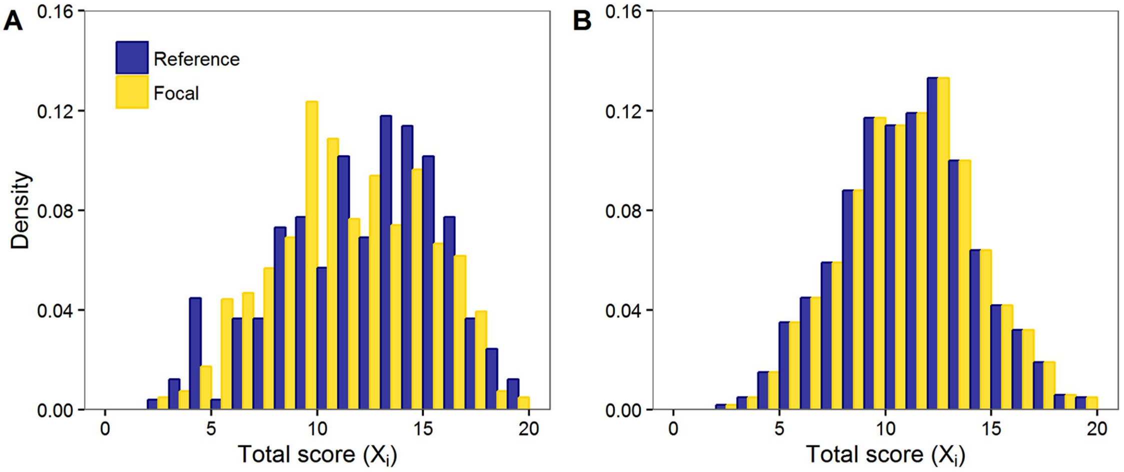

FIGURE 5. Histograms of total score by gender. Men (reference group, blue), women (focal group, yellow). (A) HCI data set. Men achieve higher scores on the HCI as confirmed by statistical tests. (B) Simulated data set based on GMAT item parameters. Distribution of total scores is exactly the same for men and women.

was incorporated into the simulation (see Supplemental Tables 5 and 6). As can be seen in Figure 6, A and B, respectively, item 1 shows uniform DIF and item 2 shows nonuniform DIF.

Comparing the Two Cases

Comparing total scores in the HCI data set suggests there is a difference between men and women in overall test performance (Table 3). However, none of the DIF methods detected any item that functioned differently for women and men. We therefore concluded that the difference in total scores was due to a real gap in how men and women understand the concepts being tested. In contrast, the analysis of the simulated data set demonstrated that, even with exactly the same distribution of total scores for both groups, the test still had items that were not functioning the same way for the two groups (Table 3). Only DIF analysis (rather than total score testing) was necessary to detect this hidden bias. If this had been a real data set, additional item analysis with content experts would be crucial for determining whether the flagged items were fair or unfair.

DIF Analysis in Biology and Beyond: A Brief Review of Other Examples

The statistical detection of DIF is only the first step in evaluating the potential measurement bias. To illustrate the necessity of content experts' review of the items tagged as DIF, we briefly describe other studies in the biology education literature in which DIF analysis has been used to identify problematic items.

In one study, Federer *et al.* (2016) explored the relationship between the way men and women answered open-ended

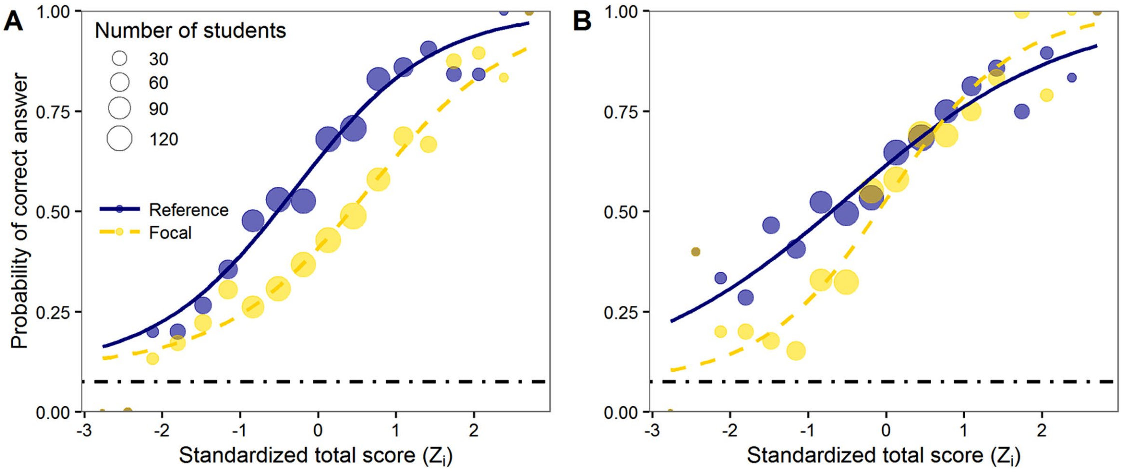

FIGURE 6. Item characteristic curves for reference (blue) and focal (yellow) group by three-parameter logistic regression model 5 for items 1 and 2 in simulated data set. Dots represent the proportion of correct answers on item by the men (reference group, blue) and women (focal group, yellow). Horizontal lines represent the estimate of the guessing parameter, *c*. (A) Uniform DIF is detected in item 1. (B) Nonuniform DIF is detected in item 2.

TABLE 3. Comparison of case studies

	HCI	Simulated data set
Total scores	Difference Men: 12.70, SD = 3.74 Women: 11.92, SD = 3.55 $p < 0.01$ (two-sample t test)	No difference Men: 11.60, SD = 3.11 Women: 11.60, SD = 3.11 $p > 0.99$ (two-sample t test)
DIF		
Mantel-Haenszel	No DIF items ($p > 0.05$ for all items)	DIF detected in item 1 ($p < 0.01$) and in item 2 ($p < 0.01$) Method does not distinguish between uniform and nonuniform DIF.
Logistic regression (model 5)	No DIF items ($p > 0.05$ for all items)	Uniform DIF detected in item 1 ($p < 0.01$) Nonuniform DIF detected in item 2 ($p < 0.01$)
IRT Wald test (model 6)	No DIF items ($p > 0.05$ for all items)	Uniform DIF detected in item 1 ($p < 0.01$) Nonuniform DIF detected in item 2 ($p < 0.01$)
Fairness	No potential unfairness detected	To be determined by content experts

questions about natural selection. They used the Mantel-Haenszel test to detect DIF and found that women performed better on questions requiring them to apply key concepts to new situations. They acknowledged that the causes of differences in gender performance are complex and need further study, while also showing evidence that the instrument they developed had little gender bias.

In another study focusing on evolution, Smith *et al.* (2016) developed an instrument to assess the extent to which high school and college students accept the theory of evolution. Their initial instrument included 14 statements in which students used a four-point rating to indicate their agreement. Two of their items were flagged as DIF, one of which was removed from the next iteration of the instrument. However, the other flagged item ("Evolution is a scientific fact.") was retained, because it helped distinguish among high school and college students. Smith *et al.*'s (2016) decision to keep one DIF item emphasizes that the statistical analysis must be paired with evaluation by content experts.

In a somewhat older study, Sudweeks and Tolman (1993) used the Mantel-Haenszel test to detect DIF and also consulted with content experts to identify potential gender-biased items for a 78-item multiple-choice test of scientific knowledge for fifth graders in Utah. The content experts found that one item was potentially biased, because they felt that one of the distractors might favor girls. However, this item was not flagged by DIF analysis. In contrast, the statistical analysis identified eight items as easier for boys and one item that was easier for girls (different than the one flagged by content experts). On the basis of these findings, the authors argued that items require both statistical and content expert analyses for developing assessments.

The consequences of unfair test items can be quite serious. Noble *et al.* (2012) responded to reports of achievement gaps in a statewide science assessment for fifth graders in Massachusetts, a form of high-stakes testing that resulted from the federal "No Child Left Behind" legislation. As part of their study, they tested DIF on a subset of six items that were flagged by experts by comparing observed item performance with content knowledge ascertained using interviews of children who took the test. Logistic regression revealed that five of the six items were indeed exhibiting DIF ($p < 0.01$). Students from low-income households and students who were English language learners were more likely to answer these items incorrectly compared with students from higher-income households or students who were native speakers—even when these focal groups had demonstrated in interviews that they correctly understood the science content. Thus, the authors concluded these test five items were unfair and needed to be revised.

In addition to being used to develop and improve instruments, DIF analysis can also be employed to study changes over time among cohorts of a population. As one example, Romine *et al.* (2016) used logistic regression to detect small differences across time in an assessment of health science interest for middle school students. Three of the items flagged for DIF revealed distinct wording differences compared with other items on the assessment. Indeed, most of the items asked students what they thought of the science they were already engaged in currently, whereas the items flagged for DIF asked students to indicate whether they wanted to spend more time learning science.

In summary, we strongly urge researchers to adopt DIF analysis as part of their routine practice in developing and improving assessments. In addition, we urge researchers to combine statistical analysis with context expertise to best understand whether DIF-flagged items are fair or unfair. This procedure helps ensure that instruments are more fair and equitable when they are first published and, further, that other research using these instruments is also more fair and equitable.

CONCLUSION

In this paper, we argue that DIF analysis is a critical part of developing both large- and small-scale educational tests, because it can be used to assess test fairness and therefore test claims about validity. Comparing the total scores of different groups is helpful to explore how different groups perform, but it is not sufficient for determining fairness. Differences in true scores might exist even in a test that is fair (case study 1). Moreover, potential unfairness of items can be hidden and not revealed by total score analysis (case study 2).

We have also provided a brief tutorial of some of the most common methods for DIF analysis, and this tutorial is supplemented with selected R code and an interactive online application (Supplemental Material; Martinková *et al.*, 2017, see also https://shiny.cs.cas.cz/ShinyItemAnalysis). As with any active area of research, more methods are available and new methods are still being proposed (e.g., Magis *et al.*, 2014; Berger and Tutz, 2016). Deciding which method to use depends on sample sizes and assumptions about items. Closer guidance may be provided by simulation studies in which data sets are generated thousands of times. This approach allows the properties of different DIF detection methods to be compared with respect to their power and type I error rate (e.g., Swaminathan and Rogers, 1990; Narayanan and Swaminathan, 1996; Güler and Penfield, 2009; Kim and Oshima, 2013; Drabinová and Martinková, 2016). Studies like these have demonstrated that the Mantel-Haenszel method works particularly well for small sample sizes but, as expected from its formula, does not always detect nonuniform DIF (Swaminathan and Rogers, 1990; Drabinová and Martinková, 2016). In our opinion, the methods that are based on regression are particularly appealing, because they are more flexible in detecting both nonuniform and uniform DIF and, unlike the Mantel-Haenszel method, also provide parameter estimates (Zumbo, 1999). Finally, IRT models have an added advantage of providing more precise estimates of latent traits, but they may be difficult to fit for sample sizes less than 500 students per group (e.g., Kim and Oshima, 2013).

We wish to reemphasize here that items flagged as DIF only have the *potential* to be unfair; an expert review is required (for methods, see, e.g., Ercikan *et al.*, 2010; Adams and Wieman, 2011) to determine whether the differences in performance among groups are due to factors related to the concept being tested, or whether they are instead “unfair” and related to a secondary latent trait, such as cultural, curricular, or language-related knowledge.

While DIF analysis is ubiquitous in large-scale assessment, it has been used rarely as a check for fairness in developing and using low-stakes tests that are used daily in all levels of education. However, developing fair tests is a value that all educators should aspire to in order to ensure that tests are not only accurate for student feedback, but also for informing modifications to teaching. Moreover, fair tests are necessary to promote and retain underrepresented groups in science, technology, engineering, and mathematics fields (Rauschenberger and Sweeder, 2010; Creech and Sweeder, 2012; Legewie and DiPrete, 2014). In short, DIF analysis should have a routine role in all our efforts to develop assessments that are more equitable measures of scientific knowledge.

ACKNOWLEDGMENTS

This research was supported by Czech Science Foundation grant number GJ15-15856Y and by National Science Foundation grant number DUE-1043443. We thank Ross H. Nehm, Roddy Theobald, Mary Pat Wenderoth, and two anonymous reviewers for their valuable feedback.

REFERENCES

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, *29*, 67–91.

Adams, W. K., & Wieman, C. E. (2011). Development and validation of instruments to measure learning of expert-like thinking. *International Journal of Science Education*, *33*, 1289–1312.

Agresti, A. (2002). *Categorical data analysis*. Hoboken, NJ: Wiley-Interscience.

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B, Statistical Methodology*, *57*, 289–300.

Berger, M., & Tutz, G. (2016). Detection of uniform and nonuniform differential item functioning by item-focused trees. *Journal of Educational and Behavioral Statistics*, *41*, 559–592.

Boone, W. J. (2016). Rasch analysis for instrument development: Why, when and how? *CBE—Life Sciences Education*, *15*, rm4.

Cai, L., Thissen, D., & du Toit, S. (2011). *IRTPRO [Software manual]. Version 2.1*, Skokie, IL: Scientific Software International. Retrieved January 24, 2016, from www.ssicentral.com/irt/index.html

Camilli, G. (2006). Test fairness. In: Brennan R. & National Council on Measurement in Education (Eds.), *Educational measurement* (pp. 220–256). Westport, CT: Praeger.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.

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

Creech, L. R., & Sweeder, R. D. (2012). Analysis of student performance in large-enrollment life science courses. *CBE—Life Sciences Education*, *11*, 386–391.

Deane, T., Nomme, K., Jeffery, E., Pollock, C., & Birol, G. (2016). Development of the Statistical Reasoning in Biology Concept Inventory (SRBCI). *CBE—Life Sciences Education*, *15*, ar5. doi: 10.1187/cbe.15-06-0131: 10.1187/cbe.15-06-0131

Dennis, J. E., Gay, D. M., Walsh, R. E., & Rice, J. (1981). An adaptive nonlinear least-squares algorithm. *ACM Transactions on Mathematical Software*, *7*, 348–368.

Doolittle, A. E. (1985). *Understanding differential item performance as a consequence of gender differences in academic background*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Downing, S. M., & Haladyna, T. M. (2006). *Handbook of test development*. Hillsdale, NJ: Lawrence Erlbaum.

Drabinová, A., & Martinková, P. (2016). Detection of differential item functioning with non-linear regression: Non-IRT approach accounting for guessing. Retrieved May 11, 2017, from http://hdl.handle.net/11104/0259498

Drabinová, A., Martinková, P., & Zvára, K. (2017). difNLR: Detection of Dichotomous Differential Item Functioning (DIF) and Differential Distractor Functioning (DDF) by Non-Linear Regression Models, R package version 1.0.0. Retrieved May 11, 2017, from https://CRAN.R-project.org/package=difNLR

Ercikan, K., Arim, R., Law, D., Domene, J., Gagnon, F., & Lacroix, S. (2010). Application of think aloud protocols for examining and confirming sources of differential item functioning identified by expert reviews. *Educational Measurement: Issues and Practice*, *29*, 24–35.

Federer, M. R., Nehm, R. H., & Pearl, D. K. (2016). Examining gender differences in written assessment tasks in biology: a case study of evolutionary explanations. *CBE—Life Sciences Education*, *15*, ar2.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.

Güler, N., & Penfield, R. D. (2009). A comparison of the logistic regression and contingency table methods for simultaneous detection of uniform and nonuniform DIF. *Journal of Educational Measurement*, *46*, 314–329.

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

Hills, J. R. (1989). Screening for potentially biased items in testing programs. *Educational Measurement: Issues and Practice*, *8*, 5–11.

Holland, P. W. (1985). *On the study of differential item performance without IRT*. In Proceedings of the 17th Annual Conference of the Military Testing Association (282–287).

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

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

Huang, X., Wilson, M., & Wang, L. (2016). Exploring plausible causes of differential item functioning in the PISA science assessment: language, curriculum or culture. *Educational Psychology*, *36*, 378–390.

IBM. (2013). *IBM SPSS statistics for Windows. Version 22.0*. Armonk, NY.

Kendhammer, L., Holme, T., & Murphy, K. (2013). Identifying differential performance in general chemistry: Differential item functioning analysis of ACS general chemistry trial tests. *Journal of Chemical Education*, *90*, 846–853.

Kim, J., & Oshima, T. C. (2013). Effect of multiple testing adjustment in differential item functioning detection. *Educational and Psychological Measurement*, *73*, 458–470.

Kingston, N., Leary, L., & Wightman, L. (1985). An exploratory study of the applicability of item response theory methods to the Graduate Management Admission Test. *ETS Research Report Series*, *1985*(2), i–56.

Legewie, J., & DiPrete, T. A. (2014). The high school environment and the gender gap in science and engineering. *Sociology of Education*, *87*, 259–280.

Libarkin, J. (2008). *Concept inventories in higher education science*. Paper presented at: National Research Council Promising Practices in Undergraduate STEM Education Workshop 2 (October 13–14, Washington, DC).

Linacre, J. M. (2005). Rasch dichotomous model vs. one-parameter logistic model. *Rasch Measurement Transactions*, *19*(3), 1032.

Liu, O. L., & Wilson, M. (2009). Gender differences in large-scale math assessments: PISA trend 2000 and 2003. *Applied Measurement in Education*, *22*, 164–184.

Lord, F. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Magis, D., Beland, S., & Raiche, G. (2016). difR: Collection of methods to detect dichotomous differential item functioning (DIF) in psychometrics. R package Version 4.7. Retrieved May 11, 2017, from https://CRAN.R-project.org/package=difR

Magis, D., Beland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods*, *42*, 847–862.

Magis, D., Tuerlinckx, F., & De Boeck, P. (2014). Detection of differential item functioning using the lasso approach. *Journal of Educational and Behavioral Statistics*, *40*, 111–135.

Mantel, M., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute: Monographs*, *22*, 719–748.

Martinello, M., & Wolf, M. K. (2012). Exploring ELL's understanding of word problems in mathematics assessments: the role of text complexity and student background knowledge. In Celedón-Pattichis S., & Ramirez N. (Eds.), *Beyond good teaching: Strategies that are imperative for English language learners in the mathematics classroom*. Reston, VA: National Council of Teachers of Mathematics.

Martinková, P., Drabinová, A., Leder, O., & Houdek, J. (2017). ShinyItemAnalysis: test and item analysis via Shiny. R package Version 1.1.0. Retrieved May 11, 2017, from https://CRAN.R-project.org/package=ShinyItemAnalysis

McFarland, J. L., Price, R. M., Wenderoth, M. P., Martinková, P., Cliff, W., Michael, J., Modell, H., & Wright, A. (2017). Development and validation of the homeostasis concept inventory. *CBE—Life Sciences Education*, *16*, ar35.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: statistical approaches for assessing measurement bias. *Applied Measurement in Education*, *17*, 297–334.

Moore, D., Notz, W., & Fligner, M. A. (2015). *The basic practice of statistics*. New York: Freeman.

Narayanan, P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement*, *20*, 257–274.

Neumann, I., Neumann, K., & Nehm, R. (2011). Evaluating instrument quality in science education: Rasch-based analyses of a nature of science test. *International Journal of Science Education*, *33*, 1373–1405. doi: 10.1080/09500693.2010.511297

Noble, T., Suarez, C., Rosebery, A., Oçonnor, M. C., Warren, B., & Hudicourt-Barnes, J. (2012). "I never thought of it as freezing": How students answer questions on large-scale science tests and what they know about science. *Journal of Research in Science Teaching*, *49*, 778–803.

Penfield, R. D., & Lee, O. (2010). Test-based accountability: potential benefits and pitfalls of science assessment with student diversity. *Journal of Research in Science Teaching*, *47*, 6–24.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*, 495–502.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, *14*, 197–207.

R Core Team (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved May 11, 2017, from www.R-project.org

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.

Rauschenberger, M. M., & Sweeder, R. D. (2010). Gender performance differences in biochemistry. *Biochemistry and Molecular Biology Education*, *38*, 380–384.

Reeves, T. D., & Marbach-Ad, G. (2016). Contemporary test validity in theory and practice: a primer for discipline-based education researchers. *CBE—Life Sciences Education*, *15*, rm1.

Romine, W. L., Miller, M. E., Knese, S. A., & Folk, W. R. (2016). Multilevel assessment of middle school students' interest in the health sciences: Development and validation of a new measurement tool. *CBE—Life Sciences Education*, *15*, ar21.

Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement*, *20*, 355–371.

Sabatini, J., Bruce, K., Steinberg, J., & Weeks, J. (2015). SARA reading components tests, rise forms: technical adequacy and test design. *ETS Research Report Series*, *2015*(2), 1–20.

SABER. (n. d.). *Biology concept inventories and assessments*. Retrieved January 24, 2016, from http://saber-biologyeducationresearch.wikispaces.com/DBER-Concept+Inventories

SAS Institute. (2013). *SAS 9.4 language reference concepts*. Cary, NC.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, *58*, 159–194.

Siegel, M. A. (2007). Striving for equitable classroom assessments for linguistic minorities: Strategies for and effects of revising life science items. *Journal of Research in Science Teaching*, *44*, 864–881.

Smith, M. U., Snyder, S. W., & Devereaux, R. S. (2016). The GAENE—Generalized Acceptance of EvolutioN Evaluation: development of a new measure of evolution acceptance. *Journal of Research in Science Teaching*, *53*, 1289–1315.

StataCorp (2015). *Stata statistical software. Release 14*. College Station, TX.

Steif, P. S., & Dantzler, J. A. (2005). A statics concept inventory: development and psychometric analysis. *Journal of Engineering Education*, *94*, 363–371.

Štuka, Č., Martinková, P., Zvára, K., & Zvárová, J. (2012). The prediction and probability for successful completion in medical study based on tests and pre-admission grades. *New Educational Review*, *28*, 138–152.

Sudweeks, R. R., & Tolman, R. R. (1993). Empirical versus subjective procedures for identifying gender differences in science test items. *Journal of Research in Science Teaching*, *30*, 3–19.

Swaminathan, H., & Rogers, J. H. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*, 361–370.

Thissen, D., Wainer, H., & Wang, X. B. (1994). Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice tests? An analysis of two tests. *Journal of Educational Measurement*, *31*, 113–123.

Walker, C. M. (2011). What's the DIF? Why differential item functioning analyses are an important part of instrument development and validation. *Journal of Psychoeducational Assessment*, *29*, 364–376.

Walker, C. M., & Beretvas, S. N. (2001). An empirical investigation demonstrating the multidimensional DIF paradigm: A cognitive explanation for DIF. *Journal of Educational Measurement*, *38*, 147–163.

Wright, C. D., Eddy, S. L., Wenderoth, M. P., Abshire, E., Blankenbiller, M., & Brownell, S. E. (2016). Cognitive difficulty and format of exams predicts gender and socioeconomic gaps in exam performance of students in introductory biology courses. *CBE—Life Sciences Education*, *15*, ar23.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). *ConQuest [computer software]*, Camberwell, Victoria: Australian Council for Educational Research.

Zenisky, A. L., Hambleton, R. K., & Robin, F. (2004). DIF detection and interpretation in large-scale science assessments: informing item writing practices. *Educational Measurement*, *9*, 61–68.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In Holland P. W. & Wainer H. (Eds.), *Differential item functioning* (pp. 337–347). Hillsdale, NJ: Erlbaum.

Zieky, M. (2003). *A DIF primer*. Princeton, NJ: Educational Testing Service. Retrieved January 24, 2016, from www.ets.org/s/praxis/pdf/dif_primer.pdf

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. Retrieved January 24, 2016, from http://faculty.educ.ubc.ca/zumbo/DIF/handbook.pdf

Zumbo, B. D. (2007). Three generations of differential item functioning (DIF) analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, *4*, 223–233.