BETTER QUALITY IN ASSESSMENTS: CONSIDERATION OF CONTEXTUAL EFFECTS ON ITEM BIAS AND DIFFERENTIAL ITEM FUNCTIONING

By

BRANDON K. VAUGHN*

* Assistant Professor, The University of Texas at Austin.

ABSTRACT

This study considers the importance of contextual effects on the quality of assessments on item bias and differential item functioning (DIF) in measurement. Often, in educational studies, students are clustered in teachers or schools, and the clusters could impact psychometric issues yet are largely ignored by traditional item analyses. A statistical model for incorporating these cluster effects will be presented. By considering DIF from this perspective, it is possible for DIF to be consistent or vary across clusters (e.g., teachers or schools). By using this model, researchers can be shown much more detail into the nature and source of DIF and provide greater quality in the analysis and use of their assessments. This can be of benefit in determining whether the nature of DIF is exclusively due to student attributes, or a particular combination of student and school attributes. In addition to this, DIF may exist only among clusters, but not for students. This is an extra detection not possible with traditional DIF analysis. For example, in some educational situations, DIF may not exist among the subjects of interest, but in certain type of schools. The use of this procedure will be demonstrated using real assessment data.

Keywords: Social Context, Cluster Effects, Multilevel Model, Differential Item Functioning, Assessments.

INTRODUCTION

The focus of this study is to consider random differential item functioning (DIF) for polytomous items from a multilevel (3 level) logistic regression perspective. Often, in educational studies, three levels with nested variables are common (e.g., items scores for students nested in schools). A statistical model for detecting random DIF for polytomously scored items will be presented.

The random-effect DIF model will incorporate a multilevel (3 levels) approach. In order to parameterize this model for polytomous outcomes, a hierarchical generalized linear model (HGLM) will be utilized. This approach will be modified to include an item response theory (IRT) model for ordinal response data. In this model, DIF may be present between any levels of the categorical response. This can be referred to as "inner-response DIF" or IDIF. In order to allow the DIF effect to randomly vary, the DIF parameters are given a random component in the level-3 model. This approach allows for the DIF effect to not be consistent across the level-3 groupings. The use of this procedure is demonstrated using real assessment data, with suggestions provided for interpreting the magnitude of the random DIF.

Background

Item bias represents a threat to the validity, and thus quality, of test scores in many different disciplines. An item is considered to be "biased" if the item unfairly favors one group over another (Buu, 2003; Holland & Wainer, 1993). More specifically, an item is considered to be biased if two conditions are met. First, performance on the item is influenced by sources other than differences on the construct of interest that are deemed to be detrimental to one group. Second, this extraneous influence results in differential performance across identifiable subgroups of examinees (Jensen, 1980).

One characteristic of bias is differential item functioning (DIF), in which examinees from different groups have differing probabilities of success on an item after being matched on the ability of interest (Borsboom, Mellenbergh, & van der Linden, 2002; Kamata & Vaughn, 2004). DIF is a necessary but insufficient condition for item bias (Williams, 1997). If an item is biased, then DIF is present. However, the presence of DIF does not imply item bias in and of itself.

An illustration of DIF is given in Figures 1 through 3. In this example, suppose there are two groups of subjects (e.g., males and females) which have different probability of a dichotomous response on an item *i*, illustrated in Figure 1. A heavier weight signifies a higher probability of getting the item correct. In Figure 1 males have a higher probability of getting this particular item correct.

Since this item is an indicator of some latent trait, then the difference between the two groups is possibly attributable to the latent trait. Therefore, controlling for this latent trait ("matching criterion") should remove the relationship between the gender and the item score. If this is the case, the item is measurement invariant across the groups. This is illustrated in Figure 2.

However, if the relationship between gender and item remains the same after controlling for the latent trait, then DIF is present. That is, the item measures something in addition to the latent trait that is differentially related to the group variable. This is shown in Figure 3.

Polytomously scored data has the additional consideration that subjects can respond to or be labeled with more than two categories on a given item. For dichotomous data, the consideration of DIF is more simplistic as there are only two outcomes. But for polytomous outcomes, there is a possibility of an "inner-response" DIF (IDIF). That is, there is the possibility that DIF may not exist uniformly across all response categories, but exist for certain responses within that item. Figure 4 illustrates an example where a particular 4-point Likert-type item displays DIF on lower ordinal responses, yet not on higher ordinal responses. This type of DIF can be referred to as a "lower" IDIF. This can exist, as an illustration, when the focal group tends to differentially vary in successfully scoring lower ordinal scores on a attitudinal measurement as compared to a reference group, while both groups have similar success in upper ordinal scoring

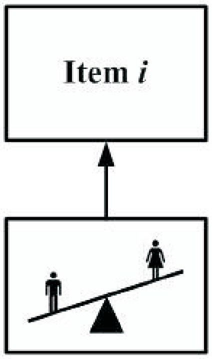

Figure 1. An illustration of gender effect

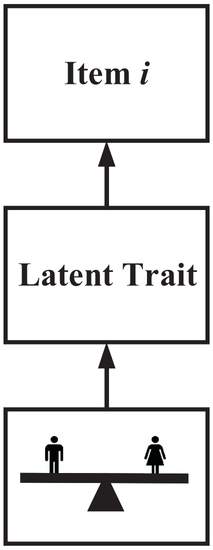

Figure 2. An illustration of no gender effect controlling for latent trait

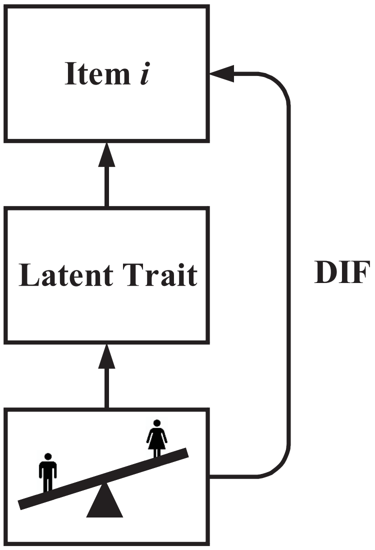

Figure 3. An illustration of DIF

categories.

Figure 5 illustrates a "balanced" IDIF, where the nature of DIF changes for both extreme ordinal responses.

In this example, there is potential bias against females on the lower ordinal responses, and potential bias against males on the upper responses. Other types of IDIF patterns are possible, For example, "upper" IDIF would indicate potential bias on the upper ordinal responses, while "consistent" IDIF would indicate that the DIF effect is approximately the same for all ordinal responses. However, patterns in IDIF are not always present, but in some situations, IDIF may only be present between certain ordinal responses and not in others with no visible pattern.

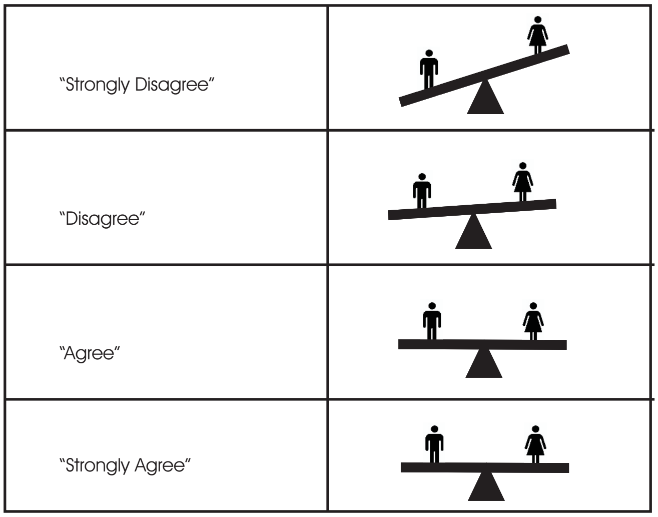

Figure 4. An illustration of lower IDIF for polytomous outcomes

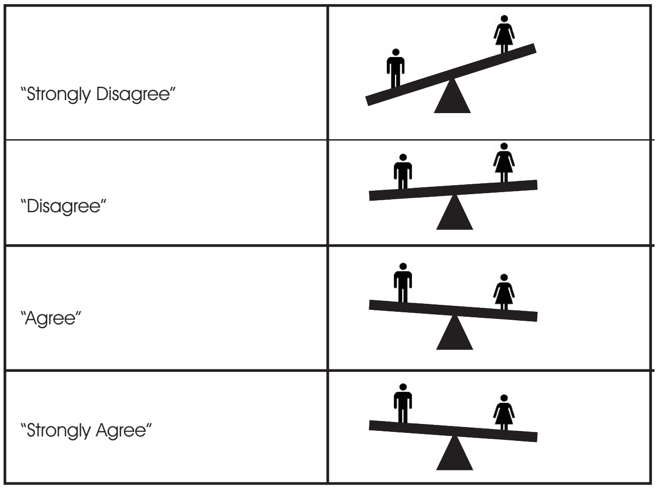

Figure 5. An illustration of balanced IDIF for polytomous outcomes

The actual assessment and measurement of DIF is not always straightforward as the concept of differential functioning. Various methods have been proposed to measure DIF. Perhaps the oldest method was an analysis of variance approach, which tested for an interaction effect between groups and items (Camilli & Shepard, 1987; Cardall & Coffman, 1964). Yet this approach did not gain in popularity due to the problematic nature of items being measured qualitatively or yielding binary outcomes (Whitmore & Schumacker, 1999). Angoff (1972) introduced one of the first widely used measures of DIF in the delta-plot method, also known as the transformed item-difficulty (TID) method. However, this method was often criticized as giving misleading results for items with differing discriminating power (Angoff, 1993; Cole, 1978; Cole & Moss, 1989; Hunter, 1975; Lord, 1977; Shepard, 1981). Various other methods were introduced, such as the Mantel-Haenszel procedure (Holland & Thayer, 1988). The Mantel-Haenszel procedure would dominate the psychometric approach to the study of DIF for many years due to its ability to give an effect size for DIF, known as α in addition to a significance test.

Another approach to DIF analysis is based on IRT principles. While traditional methods allow for items to differ in difficulty, there is no allowance for differing item discrimination (Angoff, 1993). As Angoff stresses, "it is possible for an item with the same difficulty parameter in the two groups but with different slope parameters to yield a DIF index of zero when analyzed by all but the IRT method" (p. 13). Thus, many IRT approaches to DIF emerged, in particular the multidimensional IRT method of Shealy and Stout (1993a, 1993b). The Shealy and Stout method provided an interesting vantage point for DIF analysis that the cause of DIF could be a result of multidimensionality of the test in question. One criticism of the traditional methods is that they explain very little of the source of DIF. While the IRT perspective allows for a greater discernment of DIF over traditional methods, there is still no attempt to explain the basis for DIF. One way of approaching this issue is by using multilevel analysis techniques, such as the approach proposed by Swanson, Clauser, Case, Nungster, and Featherman

(2002).

Kamata and Binici (2003) considered a multi-level approach to DIF detection. In this model, level 1 represented the item level, level 2 represented the individual level, and level 3 represented a group unit. The rationale for the inclusion of a third level was that the magnitude of DIF could vary across group units, such as schools in an educational DIF study. This approach models a random-effect DIF for the group units, and uses individual characteristics to explicate the potential sources of DIF. Chaimongkol (2005) extended the work of Kamata and Binici by using a Bayesian approach to obtain parameter estimates. The rationale for developing this approach was due to the fact that the HLM 5 software (Raudenbush, Bryk, Cheong, & Congdon, 2000) produced negatively biased parameter estimates due to a penalized quasi-likelihood (PQL) estimation method (Raudenbush, Yang, & Yosef, 2000). Raudenbush et al. and Yang (1998) suggested a sixth order Laplace (Laplace6) approximation for estimation. Current software, such as HLM 6 (Raudenbush, Bryk, & Congdon, 2005), allows for a Laplace6 approximation, but it is limited to Bernoulli models of two and three levels. Chaimongkol's work considered only a hierarchical DIF analysis for dichotomous data. Vaughn (2006) extended this work to consider polytomous outcomes.

Purpose

The primary focus of this study is to consider the social context of subjects when detecting DIF and potential item bias. Often, in educational studies, three levels with nested variables are common (e.g., item scores for students nested in schools). For each dichotomous or ordinal response, there is a probability associated with that response. Various characteristics at each level might have an effect on the response, and thus the probability of the response. These effects may be fixed or random. However, due to the fact that probability is measured on a [0,1] interval, the use of linear regression techniques is problematic. One solution is to link the linear function of these covariates to the probability of response via mathematical functions, referred to as link functions. Various link functions can be used for the modeling of dichotomous or polytomous responses, such as the logit, probit, and so on. From a traditional IRT vantage point, item effects also include discrimination and difficulty which can impact the particular response for an item. Uniform effects for all subjects are traditionally ideal, thus providing a fair and unbiased item and instrument.

One way of analyzing multilevel DIF is to focus on two levels: the first level being the item level, and the second level consisting of individual attributes. Often in these situations, DIF is considered to be as a fixed effect as shown in Figure 6. That is, any DIF effect is considered consistent across level-3 units (e.g., Schools).

An explanatory variable identifying reference/focal group affiliation can easily be added to the level-2 model and used to help measure DIF. However, the inclusion of level-3 data can have a dramatic effect on DIF estimation, especially when significant level-3 variation exists as shown in Figure 7. This variation in DIF among level-3 units is referenced as "random DIF" (RDIF). The third level will consist of grouping attributes (e.g., schools in typical educational studies). As in the case with level-2 analysis, a multi-level approach seeks to explain variation in DIF among these group attributes, as well as consider any interaction effect between the level-2 and level-3 variables.

Application

To test the effectiveness of a social context DIF model, an application of this model to real data was considered. The data set used was the NELS:88 High School Effects Study (HSES), which was sponsored by the National Center for Education Statistics (NCES). The NELS:88 focused on a

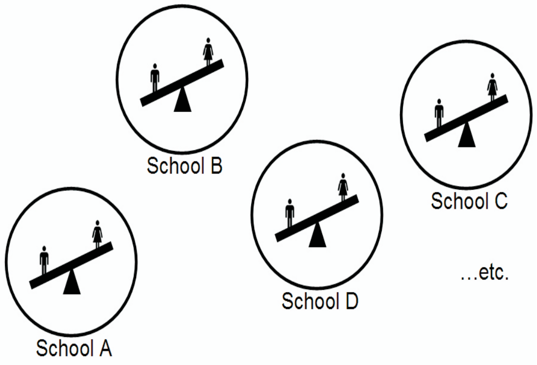

Figure 6. An illustration of a fixed DIF effect

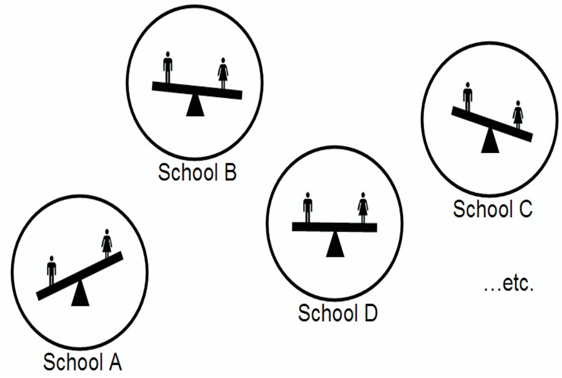

Figure 7. An illustration of a random DIF effect

national repeated measures probability sample of 10th graders (measured at the 10th through 12th grade levels) incorporating both cognitive tests and questionnaires completed by students, parents, teachers, and school administrators.

The 12th grade students completed constructed response (CR) items in the subjects of math and science. Each item was polytomously scored, as will be detailed below. Hamilton and Snow (1998) investigated DIF for gender on the science CR test of the NELS:88, and found DIF on some items using a logistic regression approach. This study also considered the science test and gender DIF, and extended the analysis to a multilevel model as presented earlier. The data consisted of records for 2190 students from 109 schools which were used to consider the effectiveness of applying the three-level random DIF model to actual research situations.

Science knowledge was assessed by four CR items. The questions required students to answer with brief written responses which might even include diagrams. The students were given 10 minutes to respond to each question. The questions were constructed so that at least part of the question could be answered by all students due to somewhat familiar content. However, only students with much deeper knowledge of the subject could answer the question in full. The four science items included:

(1) Nuclear and Fossil Fuels (CR1): Write a brief essay outlining advantages and disadvantages of each;

(2) Eclipses (CR2): Produce diagrams of solar and lunar eclipses and explain why one can be seen from a greater geographical area on earth;

(3) Rabbit and Wolf Populations (CR3): Given graph representing population of rabbits, produce graph representing population of wolves, subject to certain constraints, and explain features of graph;

(4) Heating Curve (CR4): Explain segments of graph representing temperature of a mixture as a function of time (mixture contains water and ice, and is being heated over an open flame) (Hamilton & Snow, 1998).

Readers scored each of the CR items. Most of the teams of scorers consisted of high school science teachers. The readers and test developers created a set of ordered scoring categories for each item a six-point scale score for each item, with 0 representing no understanding of the material and 5 representing complete understanding of the material. Additional information about the items and their scoring can be found in the NCES report by Rock, Pollack, and Quinn (1995). Student responses were analyzed for random DIF among gender while taking into account school affiliation. The first item was taken as the reference item since Hamilton and Snow's research indicated that the first item was relatively free of DIF. The Bayesian estimation procedures were the same as introduced in the simulation designs.

Table 1 gives the frequencies of scores at each ordinal response for each of the constructed response questions. This table is also broken down by gender, which the DIF analysis centers upon. For this analysis, females were chosen as the focal group. The frequencies at each ordinal score show relatively few students scoring at the highest levels and thus strong positive skewness.

Of special note is the difference in male and female scoring at score level 5 on question 2 (eclipses). More students were rated the highest possible score (5) on this item than any of the other three questions, yet the ratio of male to female responses is considerable.

As seen in Table 1. each question had missing data as there were a total of 2190 student records. One way to deal with this would be to use a listwise deletion method and exclude the entire subject scores if one or more

		Score						
Item		0	1	2	3	4	5	Total
Q1	Male	302	321	182	126	99	50	1080
	Female	439	340	133	82	55	25	1074
	Total	741	661	315	208	154	75	2154
Q2	Male	126	121	364	267	30	168	1076
	Female	233	236	357	174	19	50	1069
	Total	359	357	721	441	49	218	2145
Q3	Male	299	237	309	87	64	55	1051
	Female	368	269	257	81	47	37	1059
	Total	667	506	566	168	111	92	2110
Q4	Male	191	446	143	204	34	19	1037
	Female	183	416	195	209	25	15	1043
	Total	374	862	338	413	59	34	2080

Table 1. Frequencies of scoring by gender

values are missing. The method that was used in this study was replacing the missing data with a value of 0. These frequencies are presented in Table 2.

There were no students who were missing all scores for the four questions. Therefore, it was assumed in this analysis that a student missing a score for a particular question did not complete it, or possibly did not know the answer. Thus, all unanswered items were scored as incorrect answers. Although not shown in this study, a comparison in DIF analysis for this data and the data with listwise deletion showed consistent results. The dataset that was analyzed contained 2190 students from 109 schools. There were 1100 males and 1090 females. The average number of students per school was 20, with some schools reporting as few as one student while others reported a maximum of 40 students.

		Score						
Item		0	1	2	3	4	5	Total
Q1	Male	322	321	182	126	99	50	1100
	Female	455	340	133	82	55	25	1090
	Total	777	661	315	208	154	75	2190
Q2	Male	150	121	364	267	30	168	1100
	Female	254	236	357	174	19	50	1090
	Total	404	357	721	441	49	218	2190
Q3	Male	348	237	309	87	64	55	1100
	Female	399	269	257	81	47	37	1090
	Total	747	506	566	168	111	92	2190
Q4	Male	254	446	143	204	34	19	1100
	Female	230	416	195	209	25	15	1090
	Total	484	862	338	413	59	34	2190

Table 2. Frequencies of scoring by gender with missing data replacement

Parameter estimates for the random DIF model obtained using a Bayesian estimation method are presented in Table 3 and 4.

A dashed number appears beside each parameter to

Parameter	M	SD	MC error	95% CI
Item effect				
$\hat{\pi}^*_{00-1}$	-1.057	0.051	0.001	-1.158, -0.958
$\hat{\pi}^*_{00-2}$	0.522	0.047	0.001	0.430, 0.616
$\hat{\pi}^*_{00-3}$	0.535	0.047	0.001	0.443, 0.625
IDIF				
$\hat{\pi}^*_{10-1}$	0.901	0.097	0.003	0.713, 1.092
$\hat{\delta}_{210-1}$	-0.688	0.102	0.002	-0.877, -0.479
$\hat{\delta}_{220-1}$	0.737	0.159	0.004	0.388, 1.003
$\hat{\delta}_{230-1}$	0.794	0.413	0.020	-0.407, 1.154
$\hat{\delta}_{240-1}$	-0.529	0.405	0.029	-1.111, 0.988
$\hat{\pi}^*_{10-2}$	0.130	0.091	0.002	-0.047, 0.311
$\hat{\delta}_{210-2}$	-0.452	0.117	0.003	-0.671, -0.212
$\hat{\delta}_{220-2}$	0.380	0.195	0.005	-0.006, 0.748
$\hat{\delta}_{230-2}$	-0.791	0.515	0.040	-1.154, 0.934
$\hat{\delta}_{240-2}$	-0.439	0.461	0.038	-1.075, 1.109
$\hat{\pi}^*_{10-3}$	-1.031	0.104	0.003	-1.239, -0.834
$\hat{\delta}_{210-3}$	1.139	0.019	0.000	1.089, 1.155
$\hat{\delta}_{220-3}$	-1.117	0.046	0.001	-1.155, -0.997
$\hat{\delta}_{230-3}$	-0.003	0.557	0.037	-1.124, 1.029
$\hat{\delta}_{240-3}$	0.968	0.464	0.039	-0.665, 1.155
Step				
$\hat{\delta}_{200}$	1.647	0.057	0.002	1.538, 1.758
$\hat{\delta}_{300}$	1.556	0.052	0.002	1.455, 1.658
$\hat{\delta}_{400}$	1.542	0.064	0.001	1.421, 1.670
$\hat{\delta}_{500}$	0.606	0.052	0.001	0.510, 0.713
Group effect				
$\hat{\pi}^*_{010}$	0.488	0.103	0.004	0.280, 0.688

*adjusted value

Table 3. Statistics of Gibbs sampling for HSES data for fixed effects

Parameter	M	SD	MC error	95% CI
IDIF Standard Deviation				
$\hat{\sigma}_{\pi_{1-1}}$	0.634	0.359	0.007	0.030, 1.152
$\hat{\sigma}_{\pi_{2-1}}$	0.717	0.337	0.009	0.052, 1.153
$\hat{\sigma}_{\pi_{3-1}}$	0.695	0.342	0.008	0.042, 1.153
$\hat{\sigma}_{\pi_{4-1}}$	0.287	0.177	0.008	0.016, 0.657
$\hat{\sigma}_{\pi_{5-1}}$	0.720	0.379	0.023	0.038, 1.154
$\hat{\sigma}_{\pi_{1-2}}$	0.793	0.341	0.009	0.061, 1.154
$\hat{\sigma}_{\pi_{2-2}}$	0.947	0.255	0.006	0.211, 1.154
$\hat{\sigma}_{\pi_{3-2}}$	0.887	0.308	0.008	0.090, 1.154
$\hat{\sigma}_{\pi_{4-2}}$	1.037	0.160	0.010	0.568, 1.154
$\hat{\sigma}_{\pi_{5-2}}$	0.817	0.293	0.017	0.123, 1.154
$\hat{\sigma}_{\pi_{1-3}}$	0.778	0.346	0.008	0.053, 1.154
$\hat{\sigma}_{\pi_{2-3}}$	0.544	0.339	0.008	0.025, 1.147
$\hat{\sigma}_{\pi_{3-3}}$	0.625	0.359	0.010	0.033, 1.153
$\hat{\sigma}_{\pi_{4-3}}$	0.873	0.153	0.008	0.594, 1.148
$\hat{\sigma}_{\pi_{5-3}}$	0.676	0.357	0.018	0.043, 1.153
$\hat{\sigma}_{2(ref)}$	1.304	0.064	0.002	1.183, 1.434
$\hat{\sigma}_{2(focal)}$	1.047	0.067	0.003	0.916, 1.178
$\hat{\sigma}_{3}$	0.888	0.078	0.002	0.745, 1.051

Table 4. Statistics of Gibbs sampling for HSES data for random effects

help identify the question that the parameter refers to. For simplicity in notation, since the first question was used as the reference item, item 1 in the table refers to question 2, item 2 refers to question 3, and so on. The Gibbs sampler run produced 10,000 values for each parameter in the model after an initial burn-in of 1,000 iterations. The same prior distributions and initial values used in the simulation studies were used to analyze this data. Convergence using the Bayesian methods was checked, with results similar to those found in the simulation studies. The sample mean, standard deviation (SD), and MC error of these 10,000 values are presented and are interpreted in a similar manner as discussed for the DIF simulation results.

For this example, "lower IDIF" refers to DIF between ordinal scorings 0 and 1. "Upper IDIF" refers to any of the DIF estimates for the remaining comparisons. When considering the fixed DIF effects, the second and fourth question had lower IDIF values above 0.426 in magnitude. Thus, there is evidence of these two items showing DIF between the 0 and 1 ordinal scoring (those scored as having no knowledge versus little knowledge). When adding in the upper IDIF parameter estimations, the presence of DIF was also seen for some of the upper scoring as well. Although question 3 did not exhibit DIF between the first and second ordinal scoring response, there were indications of DIF on some of the upper DIF parameters. According to these upper parameters, Question 2 exhibited DIF against females, while question 3 and 4 exhibited DIF against males. When it came to the question on eclipses, possible bias against females appeared to be present. That is, for males and females with the same abilities, males seemed at an advantage of being scored in the higher ordinal categories. The opposite was true for the questions pertaining to rabbit/wolf populations and heating curves. Interestingly, the questions which gave females a possible advantage were questions dealing with graphs. These results are summarized in Table 5. The pattern of DIF is consistent with that found in the study done by Hamilton and Snow (1998).

Interpretation of Random DIF

No criteria have been established for the interpretation of random DIF. Three suggested approaches will be presented in this study. First, a quick assessment of random DIF could be seen by establishing a confidence interval for the fixed estimate. A confidence interval formed with two DIF standard deviations would give an idea of how consistent the DIF effect is. This interval is formed by calculating

If the parameter follows a normal distribution, this

$$(\textit{fixed DIF estimate}) \pm 2\hat{\sigma}_{\pi} \qquad (1)$$

	Ordinal response scoring comparison				
	0-1	1-2	2-3	3-4	4-5
Q2	? AF		? AF	? AF	? AF
Q3				? AM	? AM
Q4	? AM		? AM	? AM	

Note: AF = potential bias against females, AM = potential bias against males

Table 5. DIF results for real data study

approximates a 95% confidence interval. Regardless, based on Chebyshev's Theorem, we know that at least 75% of the estimates will fall within this interval. As illustration of this approach, a confidence interval is formed for the first DIF effect on question 2 between ordinal scores 0 and 1. The fixed DIF effect was estimated as 0.901, and the random DIF effect was estimated as 0.634. The confidence interval based on two standard deviations is (0.367, 1.268). This suggests that in some schools the DIF effect is negative, thus showing potential bias against males. Some schools may show very little DIF (values close to 0), and other schools may have a higher DIF impact against females than would be interpreted from just the fixed DIF estimate. If the confidence interval indicates a change in DIF interpretation across level-3 groupings, then an explanatory model might be warranted to explain the nature of this random DIF. If this confidence interval gives a consistent interpretation as the fixed DIF estimate (i.e., there is always potential bias toward one group across all level-3 groupings), then interpretation of the fixed DIF effect without regard to the random variation would be warranted. However, some researchers may be concerned if the variability is too high and use that as their criteria for flagging potentially biased items. In the previous example, since the nature of the DIF interpretation changes when considering the DIF variability, then this variation should be considered while interpreting the DIF effect and a follow-up explanatory analysis of level-3 groupings would be advised.

Another possible approach to the analysis of random DIF is a chi-square hypothesis test for the population variation in DIF. If the standard deviation of DIF is greater than 0.426 (the suggested value for indication of fixed DIF), then the DIF effect could become negligible or completely reverse. A possible null hypothesis for random DIF is $H_0 : \sigma^2 \leq 0.181$. The chi-square test statistic would be,

$$\chi^2 = \frac{(n-1)\hat{\sigma}^2}{0.181} \qquad (2)$$

where *n* is the level 3 sample size, and *n* 1 degrees of freedom characterize the chi-square distribution. For the previous example, the chi-square test statistic would be 419.72 which would be significant at the 5% level. However, the chi-square test is known to be sensitive to sample size and thus this approach is not recommended without the aid of the other suggestions.

A final recommended approach would be to incorporate the standard deviation of the random effect estimates from the Bayesian output to judge the magnitude in variability. This can be seen in the confidence intervals from the Gibbs sampling procedure for the posterior distribution as seen in Table 4. These are based on the posterior density for a particular parameter which may or may not be normally distributed. Any confidence interval which does not capture 0.426 and both limits are greater than 0.426 would indicate a serious variation in DIF among level 3.

A two tier approach is suggested for the analysis of random DIF. First is the assessment of random DIF:

? Negligible Random DIF: Null hypothesis (mentioned above) is rejected and the 95% posterior confidence interval captures 0.426,

? Moderate Random DIF: Null hypothesis is rejected and the 95% posterior confidence interval does not capture 0.426 but does capture 0.638,

? Large Random DIF: Null hypothesis is rejected and the 95% posterior confidence interval does not capture 0.426 or 0.638.

Second, any item which indicates at least a moderate variability in DIF can be analyzed using the first suggested method for practical importance of the DIF effect. If all items fail to reject the null hypothesis, then a fixed DIF model could be considered instead. An omnibus test for overall random DIF would be beneficial for a parsimonious decision to go with a fixed DIF model, yet development of such a test is beyond the intent of this study.

For the HSES data, all ordinal responses would be significant with the exception of the third and fourth ordinal responses on question 2 (refer back to the CR questions (1-4) presented earlier in the paper). Considering the posterior confidence intervals, most of these random DIF effect would be interpreted as negligible since they capture 0.426. The only random DIF which can be considered moderate (but not large) according to the aforementioned criteria are between ordinal scorings 3 and 4 for questions 3 and 4. The confidence interval for the DIF effect between scorings 3 and 4 for question 3 is (2.87, 1.28). The confidence interval for question 4 is (1.75, 1.74). Both of these are interpreted as indication of severe variation in the DIF effect. For question 3, the fixed DIF effect was sizeable which indicates potential bias, and inclusion of DIF variability indicates how this DIF effect varies greatly among schools. Interestingly for question 4, the fixed DIF effect was negligible; however, DIF did exist among some schools (at times in extreme amounts). This once again indicates the potential worth of a random DIF model. A fixed DIF model would have seen this particular item for these particular responses as free of DIF, when in fact there are extreme examples of DIF within the schools of this study.

One explanatory model for this random DIF could include the type of school (rural versus urban) to help explain this variability. A follow-up study in which characteristics of the schools would be incorporated would be advised at this point to investigate an explanatory model for this random DIF. These follow-up studies are not conducted in this study.

Significance

By considering DIF from a multi-level perspective, it is possible for DIF to be consistent or vary across the levels of the model. For instance, DIF may vary among level-3 groupings yet have an overall effect on level-2 subjects. For polytomous outcomes, DIF may exist for certain categorical responses consistently across level-3 groupings, or vary among these groupings. In the model for this study, level 2 represents the individual level. Therefore, most of the DIF detection will often center on this particular level. Yet with the inclusion of a third level in the model, researchers can be shown much more detail on the nature and source of DIF. Thus, one important facet of random DIF is that it allows DIF analysis to not be focused solely on the individual. Random DIF allows the social and institutional contexts in which individuals are located also to be considered.

As an illustration of this, suppose that a school district gives a survey measuring student's attitudes on a given topic. If a particular item is found to display DIF, one might like to know whether the DIF effect is consistent across all schools in the district. If it is consistent, this would be referred to as “fixed” DIF. However, if the DIF effect varies greatly among schools, then this would be an example of “random” DIF.

In a traditional educational study involving student (level 2) and school (level 3) characteristics, a multi-level approach can reveal significant interactions between the levels for DIF. This can be of benefit in determining whether the nature of DIF is exclusively due to student attributes, or a particular combination of student and school attributes. That is, certain groups of students might be affected by certain type of schools, thus causing DIF on particular items. But a traditional DIF analysis might not detect this DIF since level 3 characteristics would be aggregated into level 2 data. By considering a third level, it is possible to detect DIF for which traditional means could not. In the above example, one might find that by considering whether a school is in an urban or rural area might have a major impact on the DIF effect. Furthermore, an explanatory model could be introduced that incorporates level 3 attributes to help explain the nature of random DIF.

The uniqueness of this study is its consideration of polytomous random DIF from a multi-level perspective; it is possible for DIF to be consistent or vary across the levels of the model. For instance, DIF may vary among level 3 groupings yet have an overall effect on level 2 subjects. In addition to this, DIF may exist only for level 3 characteristics, but not level 2. This is an extra detection not possible with traditional DIF analysis. In the educational example above, in some situations DIF may not exist among all of the schools (i.e., groups of interest), but just in certain types of schools. With the inclusion of a third level in the model, researchers can be shown much more detail into the nature and source of DIF. This type of precision can be useful for test developers when the level 3 DIF seems to involve school characteristics which would be applicable to future use of the instrument. In a traditional educational study involving student (level 2) and school (level 3) characteristics, a multi-level approach can reveal significant interactions between the levels for DIF. High variability of the DIF effect could be used as part of criterion for flagging potentially biased items. For example, criteria could be established so that a certain level of variability in DIF would signal an item to be flagged as potentially biased.

References

[1]. **Angoff, W. H. (1993).** Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3-24). Hillsdale, NJ: Erlbaum.

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

[3]. **Borsboom, D., Mellenbergh, G. J., & van der Linden, W. J. (2002).** Different kinds of DIF: A distinction between absolute and relative forms of measurement invariance and bias. *Applied Psychological Measurement, 26*, 433-450.

[4]. **Buu, Y. A. (2003).** *Statistical analysis of rater effects.* Unpublished doctoral dissertation, University of Florida, Gainesville, FL.

[5]. **Camilli, G., & Shepard, L. A. (1987).** The inadequacy of ANOVA for detecting test bias. *Journal of Educational Statistics, 12*, 87-99.

[6]. **Cardall, C., & Coffman, W. E. (1964).** *A method for comparing the performance of different groups on the items of a test.* (No. Research Bulletin 64-61). Princeton, N.J: Educational Testing Service.

[7]. **Chaimongkol, S. (2005).** *Modeling differential item functioning (DIF) using multilevel logistic regression models: A Bayesian perspective.* Unpublished doctoral dissertation, Florida State University, Tallahassee, FL.

[8]. **Cole, N. S. (1978).** *Approaches to examining bias in achievement test items.* Paper presented at the national meeting of the American Personnel and Guidance Association, Washington, DC.

[9]. **Cole, N. S., & Moss, P. A. (1989).** Bias in test use. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 201-219). New York: American Council on Education/Macmillan.

[10]. **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: Lawrence Erlbaum.

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

[12]. **Hunter, J. E. (1975).** *A critical analysis of the use of item means and item-test correlations to determine the pressure or absence of content bias in achievement test items.* Paper presented at the National Institute of Education conference on test bias, Annapolis, MD.

[13]. **Jensen, A. R. (1980).** *Bias in mental testing*. New York: Free Press.

[14]. **Kamata, A., & Binici, S. (2003).** *Random-effect DIF analysis via hierarchical generalized linear models.* Paper presented at the annual meeting of the Psychometric Society, Sardinia, Italy.

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

[16]. **Lord, F. M. (1977).** A study of item bias, using item characteristic curve theory. In Y. H. Poortinga (Ed.), *Basic problems in cross-cultural psychology* (pp. 19-29). Amsterdam: Swets & Zeitlinger.

[17]. **Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., & Congdon, R. T. (2000).** HLM 5: *Hierarchical linear and nonlinear modeling [Computer Software].* Lincolnwood, IL: Scientific Software International, Inc.

[18]. **Raudenbush, S. W., Bryk, A. S., & Congdon, R. T. (2005).** HLM 6: *Hierarchical linear and nonlinear modeling (Version 6.02) [Computer software].* Lincolnwood, IL: Scientific Software International, Inc.

[19]. **Raudenbush, S. W., Yang, M. I., & Yosef, M. (2000).** Maximum likelihood for hierarchical models via high order, multivariate LaPlace approximation. *Journal of*

Computational and Graphical Statistics, 9(1), 141-157.

[20]. **Raudenbush, S. W., Yang, M.-L., & Yosef, M. (2000).** Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *Journal of Computational and Graphical Statistics, 9*, 141-157.

[21]. **Rock, D. A., Pollack, J. M., & Quinn, P. (1995).** *Psychometric Report for the NELS:88 Base Year through Second Follow-Up. National Education Longitudinal Study of 1988.* (No. NCES-95-382). Washington, DC: National Center for Education Statistics.

[22]. **Shealy, R. T., & Stout, W. F. (1993b).** An item response theory model for test bias and differential item functioning. In P. Holland & H. Wainer (Eds.), *Differential item functioning*. Hillsdale, NJ: Erlbaum.

[23]. **Shealy, R. T., & Stout, W. F. (1993a).** 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.

[24]. **Shepard, L. A. (1981).** Identifying bias in test items. In B. F. Green (Ed.), *New direction in testing and measurement: Issues in testing, coaching, disclosure and test bias* (Vol. 11, pp. 79-104). San Francisco: Jossey-Bass.

[25]. **Swanson, D. B., Clauser, B. E., Case, S. M., Nungster, R. J., & Featherman, C. (2002).** Analysis of differential item functioning (DIF) using hierarchical logistic regression models. *Journal of Educational and Behavioral Statistics, 27*, 53-75.

[26]. **Vaughn, B. K. (2006).** *A Hierarchical Generalized Linear Model of Random Differential Item Functioning for Polytomous Items: A Bayesian Multilevel Approach.* Unpublished doctoral dissertation, Florida State University, Tallahassee, FL.

[27]. **Whitmore, M. L., & Schumacker, R. E. (1999).** A comparison of logistic regression and analysis of variance differential item functioning detection methods. *Educational and Psychological Measurement, 59*, 910-927.

[28]. **Williams, V. S. L. (1997).** The "unbiased" anchor: Bridging the gap between DIF and item bias. *Applied Measurement in Education, 10*, 253-267.

[29]. **Yang, M. L. (1988).** *Increasing the efficiency in estimating multilevel Bernoulli models.* Unpublished doctoral dissertation, Michigan State University, East Lansing, MI.

ABOUT THE AUTHOR

Brandon Vaughn's current research interests include: multi-level differential item functioning (DIF), Bayesian estimation procedures, creative uses of non-parametric classification procedures, and effective strategies in the teaching of statistics.