Doc 4.9

Comparability of English and Spanish/English Mathematics Tests: Differential Item Functioning

Oregon Department of Education
October 2010

Introduction

Issues of equity and fairness have long been concerns of the broader society (Camilli, 2006), while the related matters associated with test and item bias have been of particular interest to psychometric community when testing the mathematics skills of limited English speakers. For test results to be considered comparable, the conditions of testing should be "reasonable and equitable", thereby generalizing to all subpopulations taking the test. For this reason, language accommodations like language glossaries and side-by-side translations are often utilized in an effort to limit any "construct irrelevant" variation in the scores. Employing such tactics, Oregon has untaken efforts to produce meaningful accommodations designed to provide such linguistic alternatives in an effort to increase score comparability. Despite these efforts, to understand whether such accommodations produce any meaningful remedy, the 1999 Standards for Educational and Psychological Testing demands that one examines whether the employed items and resulting scores are equally valid given the purposes of the assessment.

If a test item is equitable, it is administered without any advantage being provided to any one group or class of individuals. None-the-less, some "protected" groups have historically experienced previous forms of discrimination and compensatory remedies have been applied in an effort to provide more equity. A group or class is identified as having protected status when previous government or civil actions potentially impacted their rights. Historically, the Supreme Court has awarded protected status to a number of groups covering race, national origin, gender, and disability. Because of this protected status, we actively research ways of maintaining test comparability between these protected groups and other subgroups. Many students who belong to these protected groups have impediments that are not relevant to the construct being measured and interfere with the valid assessment of knowledge and skills. In these cases, previously approved accommodations may be made available so that a more accurate measure of the student's proficiencies may be obtained.

An accommodation represents some departure from the standardized administrative practice made in response to some student need but does not change the construct validity of inferences made with the assessment (Koretz & Hamilton, 2006). In Oregon, test accommodations are changes or adjustments in the administrative practices of the test undertaken to increase the accuracy and construct relevance of the scores. In theory, these changes or adjustments in test administration are developed and implemented in ways **not designed** to affect the proficiencies attributed to the constructs being measured. So, for example, to reduce the chance of any advantage to any language subgroup, test administrators can read the mathematics test to the student or provide a side-by-side translation of the test in the student's primary language. The mathematical concepts and procedures measured by the test are separate from the student's language or reading abilities, so the intent is to reduce the construct irrelevant variance attributed to the students' verbal abilities. However, because a reading test intentionally measures reading and language abilities, the test administrator does not permit reading accommodations for any validly administered reading test.

The primary purpose of the side-by-side accommodation is to provide a comparable score that is less sensitive to linguistic differences, thereby improving the validity of the score. The intent of the side-by-side accommodation is to reduce any language load so that the limited English Speaker or focal group better understands what is being asked without affecting the item's difficulty. Reasonable changes and adjustments to administrative practices provide many students with a more equal opportunity to demonstrate their true level of knowledge and skills, reducing the chance of any adverse impact. However, any linguistic adjustment potentially has some deleterious effects when linguistic adjustments cue solutions and affect item difficulty. So, for example, a Spanish language translation that makes the item easier provides an unintended and systematic advantage to the protected group.

Some invariance in items measuring mathematical performance is observed between subgroups when changes occur in curriculum expectations. For whatever the reasons, such differences are often observed when one subgroup has not had the opportunity to learn the concepts. Performance differences attributed to a lack of opportunity to learn the concepts are not remedied by dropping or rewriting targeted items. Such disparate impacts are appropriately remedied by teaching the material to the underachieving group.

The purpose of the current research is to identify both uniform and non-uniform differences in mathematical performance across a number of ability levels for groups of students taking the side-by-side test and those taking the standard mathematics test in English. With these purposes in mind, the following research questions are addressed:

> Do side-by-side items in mathematics written and administered in both English and Spanish provide a fair test of a limited English speaker's mathematics achievement?
>
> Do these accommodations in administrative practice provide any unintended advantages to any one subgroup?

As an additional benefit, after identifying any item with differential item functioning, content experts might use the results to minimize any unintended differences in the performance attributed to these language differences.

Differential Item Functioning Study

One technical approach for empirically examining the fairness of each item involves estimating whether the probability of success for each linguistic group is similar at each ability level. A Differential Item Function (DIF) Study tests for between group differences in item performance across alternative levels of the ability distribution. Limited English speakers who have protected status and are taking the side-by-side accommodation are referred to as the "focal" group in this DIF study; English speakers without protected status are referred to as the "referent" group in a DIF study.

Most DIF studies potentially examine both uniform and non-uniform differences in the success rates of two or more groups on an item at a given ability level. Uniform differences describe constant difference of one group's success rates above or below another group's rates. Non-uniform differences describe alternating differences in success rates at different levels of the score distribution for each linguistic group. Large DIF effects are necessary conditions when evaluating potential item bias, but group differences in item functioning **do not always imply that the item is biased**. Competing explanations can often occur that better explain these differences, especially when similar patterns exist across various items within a specific content area. Judgmental or logical analysis must then be used to make an ethical decision regarding the future use of the item.

Because Oregon's Knowledge and Skills tests are adaptive, most students are taking a random form of a tailored test that is generated on the fly and weighted by the content standards. Since each test is a fairly unique compilation of forty or more items that is adapted to the person's ability level, the person's raw score is less suitable as a matching variable. For any DIF analysis using adaptive scores, the person's RIT score is employed as the conditioning variable to match the various abilities of persons within each of the subgroups.

Statistical differences in the estimated difficulty estimates of the items at a given ability level is assumed to represent some form of systematic error that potentially produces construct irrelevant variance or bias (Camilli, 2006). In this case, a mathematics item demonstrating DIF is language biased against the protected group when more proficiency in the primary language is necessary for the limited English speaker to demonstrate his/her true understanding of the mathematics necessary for solving a problem. Differential difficulties are often not obvious since some linguistic differences may be attributed to language disabilities, opportunities to learn, or cultural differences associated with understanding one's primary language.

Sampling

According to Zwick (2000), to form comparable groups, one needs:

1. An appropriate matching variable.
2. A match with stable results in small samples.

When analyzing adaptive tests, Stiennberg, *et al*. (1990) suggested matching on the true score and applying DIF methods. To obtain comparable groups, a sampling program first matches the distribution of the scores of students in the referent group to the existing distribution of scores of students in the focal group. The program then segments the focal group's distribution of scores into several intervals, and then randomly selects students from the reference group with scores that match the students' scores in the focal group within each interval. So, for example, if 5% of the students in the focal group had scored between 200 and 210 on the test, the sampling program would match the scores of students in referent group until 5% of those scores were between 200 and 210. This matching procedure is performed across the entire distribution of scores in the focal group until a similar distribution of matched scores was generated for the reference group. If sufficient numbers of students were available at each ability level, the item means and standard deviations were approximately equal for both the focal and reference groups after sampling.

ODE attempted to generate approximately equal numbers of students for each analysis, but analysts limited the size of the focal group to 500 students to best achieve reasonable matches at all ability levels. A sufficiently large sample size is assumed, as in all significance testing. The Mantel-Haenszel procedure has been found to more applicable than most DIF methods, but even this procedure has its limits. Mazor, Clauser, and Hambleton (1992) suggested that samples smaller than 100 in the reference and focal groups were too small. They recommended samples as large as 200 per group to make adequate decisions. Valid results with relatively small samples are an advantage seen in these sampling procedures. A sample too small would produce abnormally large Type II error rates, allowing items with DIF to go undetected when there was, in fact, a real difference in the probability of the correct response between groups. For this reason, sufficient sample size for judging DIF with these matching methods was determined to be about 250 per group. This strategy also permitted most ability levels summarized in cells to have 5 or more cases.

Methods Employed to Study DIF

Mantel-Haenszel Procedure: Holland (1985) proposed the use of the Mantel-Haneszel procedure as a practical and powerful way to detect test items that function differently for two matched groups of examinees. A 2x2 cross-tabulation table is produced for the previously matched set of examinees in both the reference and focal groups over each of the K levels of ability. The Mantel-Haenszel procedure tests the null hypothesis that the common odds ratio of correct response across all matched groups is α = 1 over the K levels. Mantel and Haenszel developed an estimator of α whose scale ranges from 0 to ∞ known as alpha ($\hat{\alpha}$), so an obtained value of $\hat{\alpha}$=1 implies that there is negligible or no DIF. A small, obtained value less than 1 favors the focal group, while a large value greater than 1 favors the referent group.

Item Scoring

Group	Correct=1	Incorrect=0	Total
Reference (R)	A $_1$	B $_1$	n $_{R1}$
Focal (F)	C $_1$	D $_1$	n $_{f1}$
Total	m $_{11}$	m $_{00}$	T $_1$
Reference (R)	A $_2$	B $_2$	n $_{R2}$
Focal (F)	C $_2$	D $_2$	n $_{f2}$
Total	m $_{12}$	m $_{02}$	T $_2$
Reference (R)	A $_{K-1}$	B $_{K-1}$	n $_{RK-1}$
Focal (F)	C $_{K-1}$	D $_{K-1}$	n $_{fK-1}$
Total	m $_{1K-1}$	m $_{0K-1}$	T $_{K-1}$
Reference (R)	A $_K$	B $_K$	n $_{RK}$
Focal (F)	C $_K$	D $_K$	n $_{fK}$
Total	m $_{1K}$	m $_{0K}$	T $_K$

$$\hat{\alpha}_{MH} = \frac{\sum_K A_K D_K / T_K}{\sum_K B_K C_K / T_K}$$

Since alpha is not symmetric, Holland and Thayer (1985) proposed a natural log transformation of the Mantel and Haenszel's estimator called "delta" that is symmetric and has 0 as a null value. A delta value close to or equal to 0 has no DIF, a negative delta value significantly less than 0 corresponds to items the reference group found easier to get correct, and delta values significantly greater than 0 corresponds to items the focal group found easier to get correct.

$$\text{MH D-DIF= } \Delta_{MH} = -2.35 * \ln(\hat{\alpha}_{MH})$$

The absolute value of delta is related to the ETS items scale of item difficulty called the delta plot scale. The delta plot method calls for the calculation of p-values for both groups being examined and conversion of each p-value to a normal deviate scale that has a mean of 13 and a standard deviation of 4 (see Crocker & Algina, 1986 for details). Values of the delta plot scale for every item are also found in the classical statistics report received by content specialists.

Using a classification system defined by Zieky (1993) at ETS, when the absolute value of delta d-dif has a magnitude that:

- Equals or exceeds 1.0 and is less than 1.5 in absolute units, with a MH chi square value that is significant (p<.05), the item is assigned a **B** rating. Type **B** items are considered to have moderate DIF and are commonly retained on the test.

- Equals or exceeds 1.5 in absolute units, with a MH chi square value is statistically significant (p<.05), the item is assigned a **C** rating. Type **C** items are considered to have DIF with the most magnitude and should be considered for removal.

- Provides any other test combination of test result with neither a likelihood chi-square value that is significantly different from 0 nor a d-diff value greater than 1 is classified with an **A** rating. An item classified as Type **A** demonstrates no evidence of DIF.

Logistic Regression: The Mantel-Haenszel method assumes that only the difficulty of the items may change, and item DIF is detected by simultaneously testing for significant group differences in the odds ratio at K ability levels. Like the Mantel-Haenszel procedure, the logistic regression first tests for "uniform" differences in the responses represented by comparing the fit of the model. This is done by first fitting a model relating the dichotomous response of the item to the RIT score and calculating a chi-square value. A second model expands on the first model by adding a group variable to the original model and using the likelihood ratio test (1 df) to examine changes in the fitted model. By subtracting the chi-square value of the second model from the first, a likelihood chi-square test of difference is calculated. A significant change in model fit means there is significant uniform DIF. This approach has been shown to be mathematically comparable to the Mantel Haenszel result.

Unlike the more restrictive Mantel Haenszel method, logistic regression goes further by testing whether the item discriminates equally well across the ability distribution. Items with non-uniform DIF identify groups who have advantage over a second group in one area of the distribution, but are at a disadvantage at another end of the distribution. One way to test for such differences in the rates of growth between groups is to fit a model that adds an interaction term to the second model. This third model with its RIT term, its group term, and its RIT by group interaction term is fitted and a change in the chi-square value is calculated by subtracting the chi-square value of the third model from the second model. Any significant change in chi-square would suggest non-uniform DIF.

The three models to be compared are:

Model 1 $z = \beta_0 + \beta_1 X$

Model 2 $z = \beta_0 + \beta_1 X + \beta_2 G$

Model 3 $z = \beta_0 + \beta_1 X + \beta_2 G + \beta_3 XG$

where $z = \ln(P/Q)$

$Q=1-P$

Likelihood Difference Test

$G = \chi^2$ = D(For the model without the variable) – D(For the model with the variable)

$= D_{model1} - D_{model2}$

$= D_{model2} - D_{model3}$

Like the Mantel Haenszel's procedure, the size of the DIF in a logistic regression analysis has been studied and classified by Jodoin and Gierl (1999).

- Negligible or No DIF-level DIF: $R^2 \Delta$-U < 0.035,
- Moderate level DIF: Null hypothesis is rejected and $0.035 \leq R^2 \Delta$-U < 0.070,
- Large level DIF: Null hypothesis is rejected and $R^2 \Delta$ -U $\geq$ 0.070.

The Jodoin and Gierl research further suggested that Type I error rates in DIF analysis were often inflated. As a result, a more conservative set of classifications were adopted by these researchers. As a basis of comparison, the Oregon Department of Education suggests a "small" DIF class between the moderate and No DIF classes of the Jodoin/Gierl classifications (Small level DIF: $0.020 \leq R^2 \Delta$-U < 0.035), so that a more complete analysis of the differences in the Mantel Haenszel and Jodoin/Gierl classification systems can be noted within the context of this study.

Results of the Study

Although the number of students taking most items in the grade-level pools was low, the results of the Mantel Haenszel and logistic regression analysis demonstrated little or no DIF for the preponderance of items utilized within at all the grade level pools. Summaries of these results may be found in Tables 1-14. For the Mantel Haenszel Tables 1-7, between 86 -90 percent of the items demonstrated No DIF and were classified A in each pool, another 5-7 percent had moderate levels of DIF and were classified as B items, while a final 5-7 percent of the items had large levels of DIF and were classified as C items. Considering the number of items and the number of tests of DIF, the probability of identifying an item with DIF was just above chance levels. In other words, given an alpha level =0.05 and considering the hundreds of statistical tests performed, there is a 5% possibility of falsely rejecting the null by chance alone. Given that the nominal alpha and the actual or exact alpha are so close after so many tests, it is difficult to determine what items are problematic without additional tests over time or across groups. However, this assumes a random sample of students taking the each item.

One pattern is apparent when examining the number of items with significance tests classified as having large or C DIF. As sample size increases, with the exception of grade 8, the number of items classified as large DIF or C type items with 200 or more persons in a group decreases compared to the B type items. However, samples of 200 in each group when analyzing side-by-side items are still smaller than the 500 group size typically employed for DIF analysis in the standard item pools. DIF analysis based on restricted samples of responses that are applied by contingency table approaches like Mantel Haenszel have been shown to be limited (Fidalgo, Ferreres, & Muniz, 2004). Conservatively, these authors recommend using higher significance levels (*=.20) as opposed to ETS classification systems designed with effect size in mind. Conservatively, any type C items with 200 or more persons in the focal group should be examined by specialists and considered for removal from the pools. In addition, when groups are smaller than 200 cases and one applies the Fidalgo et. al. recommendations, more DIF items are identified than found using the ETS classification system.

Logistic Regression provides another look at DIF using the same group performances. Jodoin/Gierl effect size criteria are applied to classify the previous DIF results as either moderate or large. Results indicate the many of the same items are identified using logistic regression, but many items classified as C type items with Mantel Haenszel statistics indicating large DIF are now reclassified as having moderate or even small DIF using our addition to the Jodoin/Gierl effect classification system. One possible explanation is that, once the interaction effects are included in the DIF model, the uniform DIF effects are dampened and their statistical effects become less severe. This is analogous to what happens in a factorial ANOVA design or in the linear regression when the interaction effects are included in the model. --uniform or main effects appear to dampen and their significance level falls. This result could account for differences in the Type I rates found when comparing the Jodoin/Gierl and ETS classification systems. However, there may be an alternative explanation for the low numbers of large DIF items identified using logistic regression. The small and restricted samples may possess less power to detect effects when they potentially exist. Under these conditions, increased Type II error is more of a potential explanation given that Jodoin/Gierl used samples of 250 more in each group to establish their recommended classification system. Tables 15 to 21 summarize these declines in large DIF effects when applying the Jodoin/Gierl classifications for each pool at every grade level.

Appendix A lists all the item results for both the Mantel Haenszel and logistic regression analyses. It is possible that one could use the p-value of 0.20 on the chi square test to apply a conservative approach to the examination of the DIF results.

Mantel-Haenszel Statistics Summary

Focal Group Responses	Counts and Percents	ETS MH DIF Classification			Total
		A	B	C	
Below 100 Cases	Count	104	0	6	110
	% within Group Size	94.5%	.0%	5.5%	100.0%
	% within Classification	18.2%	.0%	18.8%	17.1%
Between 100 and 200 Cases	Count	206	8	18	232
	% within Group Size	88.8%	3.4%	7.8%	100.0%
	% within Classification	36.1%	19.0%	56.3%	36.0%
Over 200 Cases	Count	261	34	8	303
	% within Group Size	86.1%	11.2%	2.6%	100.0%
	% within Classification	45.7%	81.0%	25.0%	47.0%
Total	Count	571	42	32	645
	% within Group Size	88.5%	6.5%	5.0%	100.0%
	% within Classification	100.0%	100.0%	100.0%	100.0%

Table 1
Grade 3 Items

Focal Group Responses	Counts and Percents	ETS MH DIF Classification			Total
		A	B	C	
Below 100 Cases	Count	123	0	6	129
	% within Group Size	95.3%	.0%	4.7%	100.0%
	% within Classification	24.5%	.0%	15.8%	22.2%
Between 100 and 200 Cases	Count	142	7	18	167
	% within Group Size	85.0%	4.2%	10.8%	100.0%
	% within Classification	28.2%	17.5%	47.4%	28.7%
Over 200 Cases	Count	238	33	14	285
	% within Group Size	83.5%	11.6%	4.9%	100.0%
	% within Classification	47.3%	82.5%	36.8%	49.1%
Total Items	Count	503	40	38	581
	% within Group Size	86.6%	6.9%	6.5%	100.0%
	% within Classification	100.0%	100.0%	100.0%	100.0%

Table 2
Grade 4 Items

Focal Group Responses	Counts and Percents	ETS MH DIF Classification			Total
		A	B	C	
Below 100 Cases	Count	130	0	3	133
	% within Group Size	97.7%	.0%	2.3%	100.0%
	% within Classification	25.9%	.0%	10.7%	23.5%
Between 100 and 200 Cases	Count	185	10	20	215
	% within Group Size	86.0%	4.7%	9.3%	100.0%
	% within Classification	36.9%	27.0%	71.4%	37.9%
Over 200 Cases	Count	187	27	5	219
	% within Group Size	85.4%	12.3%	2.3%	100.0%
	% within Classification	37.3%	73.0%	17.9%	38.6%
Total Items	Count	502	37	28	567
	% within Group Size	88.5%	6.5%	4.9%	100.0%
	% within Classification	100.0%	100.0%	100.0%	100.0%

Table 3
Grade 5 Items

Focal Group Responses	Counts and Percents	ETS MH DIF Classification			Total
		A	B	C	
Below 100 Cases	Count	165	0	16	181
	% within Group Size	91.2%	.0%	8.8%	100.0%
	% within Classification	33.7%	.0%	45.7%	33.5%
Between 100 and 200 Cases	Count	236	8	15	259
	% within Group Size	91.1%	3.1%	5.8%	100.0%
	% within Classification	48.3%	50.0%	42.9%	48.0%
Over 200 Cases	Count	88	8	4	100
	% within Group Size	88.0%	8.0%	4.0%	100.0%
	% within Classification	18.0%	50.0%	11.4%	18.5%
Total Items	Count	489	16	35	540
	% within Group Size	90.6%	3.0%	6.5%	100.0%
	% within Classification	100.0%	100.0%	100.0%	100.0%

Table 4
Grade 6 Items

Focal Group Responses	Counts and Percents	ETS DIF Classification			Total
		A	B	C	
Below 100 Cases	Count	187	0	6	193
	% within Group Size	96.9%	.0%	3.1%	100.0%
	% within Classification	50.5%	.0%	26.1%	46.7%
Between 100 and 200 Cases	Count	99	2	9	110
	% within Group Size	90.0%	1.8%	8.2%	100.0%
	% within Classification	26.8%	10.0%	39.1%	26.6%
Over 200 Cases	Count	84	18	8	110
	% within Group Size	76.4%	16.4%	7.3%	100.0%
	% within Classification	22.7%	90.0%	34.8%	26.6%
Total Items	Count	370	20	23	413
	% within Group Size	89.6%	4.8%	5.6%	100.0%
	% within Classification	100.0%	100.0%	100.0%	100.0%

Table 5
Grade 7 Items

Focal Group Responses	Counts and Percents	ETS DIF Classification			Total
		A	B	C	
Below 100 Cases	Count	176	0	9	185
	% within Group Size	95.1%	.0%	4.9%	100.0%
	% within Classification	43.3%	.0%	37.5%	41.4%
Between 100 and 200 Cases	Count	167	6	8	181
	% within Group Size	92.3%	3.3%	4.4%	100.0%
	% within Classification	41.1%	35.3%	33.3%	40.5%
Over 200 Cases	Count	63	11	7	81
	% within Group Size	77.8%	13.6%	8.6%	100.0%
	% within Classification	15.5%	64.7%	29.2%	18.1%
Total Items	Count	406	17	24	447
	% within Group Size	90.8%	3.8%	5.4%	100.0%
	% within Classification	100.0%	100.0%	100.0%	100.0%

Table 6
Grade 8 Items

Focal Group Responses	**Counts and Percents**	**ETS DIF Classification**			Total
		A	**B**	**C**	
Below 100 Cases	Count	533	0	16	549
	% within Group Size	97.1%	.0%	2.9%	100.0%
	% within Classification	80.4%	.0%	57.1%	79.0%
Between 100 and 200 Cases	Count	122	3	12	137
	% within Group Size	89.1%	2.2%	8.8%	100.0%
	% within Classification	18.4%	75.0%	42.9%	19.7%
Over 200 Cases	Count	8	1	0	9
	% within Group Size	88.9%	11.1%	.0%	100.0%
	% within Classification	1.2%	25.0%	.0%	1.3%
Total Items	Count	663	4	28	695
	% within Group Size	95.4%	.6%	4.0%	100.0%
	% within Classification	100.0%	100.0%	100.0%	100.0%

Table 7
Grade 10 Items

Logistic Regression Statistics Summary

Focal Group Responses	Counts and Percents	Uniform Effect Size				Total
		No DIF	Large	Moderate	Small	
Below 100 Cases	Count	81	3	9	16	109
	% within Group Size	74.3%	2.8%	8.3%	14.7%	100.0%
	% within Classification	14.1%	100.0%	52.9%	34.8%	17.0%
Between 100 and 200 Cases	Count	208	0	6	17	231
	% within Group Size	90.0%	.0%	2.6%	7.4%	100.0%
	% within Classification	36.2%	.0%	35.3%	37.0%	36.1%
Over 200 Cases	Count	285	0	2	13	300
	% within Group Size	95.0%	.0%	.7%	4.3%	100.0%
	% within Classification	49.7%	.0%	11.8%	28.3%	46.9%
Total Items	Count	574	3	17	46	640
	% within Group Size	89.7%	.5%	2.7%	7.2%	100.0%
	% within Classification	100.0%	100.0%	100.0%	100.0%	100.0%

Focal Group Responses	Counts and Percents	Interaction Effect Size				Total
		No DIF	Large	Moderate	Small	
Below 100 Cases	Count	88	4	5	12	109
	% within Group Size	80.7%	3.7%	4.6%	11.0%	100.0%
	% within Classification	14.5%	100.0%	100.0%	50.0%	17.0%
Between 100 and 200 Cases	Count	220	0	0	11	231
	% within Group Size	95.2%	.0%	.0%	4.8%	100.0%
	% within Classification	36.2%	.0%	.0%	45.8%	36.1%
Over 200 Cases	Count	299	0	0	1	300
	% within Group Size	99.7%	.0%	.0%	.3%	100.0%
	% within Classification	49.3%	.0%	.0%	4.2%	46.9%
Total Items	Count	607	4	5	24	640
	% within Group Size	94.8%	.6%	.8%	3.8%	100.0%
	% within Classification	100.0%	100.0%	100.0%	100.0%	100.0%

Table 8
Grade 3 Items

Focal Group Responses	Counts and Percents	Uniform Effect Size				Total
		No DIF	Large	Moderate	Small	
Below 100 Cases	Count	93	3	19	14	129
	% within Group Size	72.1%	2.3%	14.7%	10.9%	100.0%
	% within Classification	18.4%	75.0%	70.4%	31.1%	22.2%
Between 100 and 200 Cases	Count	143	1	2	21	167
	% within Group Size	85.6%	.6%	1.2%	12.6%	100.0%
	% within Classification	28.3%	25.0%	7.4%	46.7%	28.7%
Over 200 Cases	Count	269	0	6	10	285
	% within Group Size	94.4%	.0%	2.1%	3.5%	100.0%
	% within Classification	53.3%	.0%	22.2%	22.2%	49.1%
Total Items	Count	505	4	27	45	581
	% within Group Size	86.9%	.7%	4.6%	7.7%	100.0%
	% within Classification	100.0%	100.0%	100.0%	100.0%	100.0%

Focal Group Responses	Counts and Percents	Interaction Effect Size				Total
		No DIF	Large	Moderate	Small	
Below 100 Cases	Count	112	2	6	9	129
	% within Group Size	86.8%	1.6%	4.7%	7.0%	100.0%
	% within Classification	20.1%	100.0%	100.0%	56.3%	22.2%
Between 100 and 200 Cases	Count	161	0	0	6	167
	% within Group Size	96.4%	.0%	.0%	3.6%	100.0%
	% within Classification	28.9%	.0%	.0%	37.5%	28.7%
Over 200 Cases	Count	284	0	0	1	285
	% within Group Size	99.6%	.0%	.0%	.4%	100.0%
	% within Classification	51.0%	.0%	.0%	6.3%	49.1%
Total Items	Count	557	2	6	16	581
	% within Group Size	95.9%	.3%	1.0%	2.8%	100.0%
	% within Classification	100.0%	100.0%	100.0%	100.0%	100.0%

Table 9
Grade 4 Items

Focal Group Responses	Counts and Percents	Uniform Effect Size				Total
		No DIF	Large	Moderate	Small	
Below 100 Cases	Count	95	4	11	23	133
	% within Group Size	71.4%	3.0%	8.3%	17.3%	100.0%
	% within Classification	19.3%	80.0%	57.9%	45.1%	23.5%
Between 100 and 200 Cases	Count	188	0	7	20	215
	% within Group Size	87.4%	.0%	3.3%	9.3%	100.0%
	% within Classification	38.2%	.0%	36.8%	39.2%	37.9%
Over 200 Cases	Count	209	1	1	8	219
	% within Group Size	95.4%	.5%	.5%	3.7%	100.0%
	% within Classification	42.5%	20.0%	5.3%	15.7%	38.6%
Total Items	Count	492	5	19	51	567
	% within Group Size	86.8%	.9%	3.4%	9.0%	100.0%
	% within Classification	100.0%	100.0%	100.0%	100.0%	100.0%

Focal Group Responses	Counts and Percents	Interaction Effect Size				Total
		No DIF	Large	Moderate	Small	
Below 100 Cases	Count	105	3	12	13	133
	% within Group Size	78.9%	2.3%	9.0%	9.8%	100.0%
	% within Classification	19.7%	100.0%	80.0%	76.5%	23.5%
Between 100 and 200 Cases	Count	209	0	3	3	215
	% within Group Size	97.2%	.0%	1.4%	1.4%	100.0%
	% within Classification	39.3%	.0%	20.0%	17.6%	37.9%
Over 200 Cases	Count	218	0	0	1	219
	% within Group Size	99.5%	.0%	.0%	.5%	100.0%
	% within Classification	41.0%	.0%	.0%	5.9%	38.6%
Total Items	Count	532	3	15	17	567
	% within Group Size	93.8%	.5%	2.6%	3.0%	100.0%
	% within Classification	100.0%	100.0%	100.0%	100.0%	100.0%

Table 10
Grade 5 Items

Focal Group Responses	Counts and Percents	Uniform Effect Size				Total
		No DIF	Large	Moderate	Small	
Below 100 Cases	Count	144	7	15	15	181
	% within Group Size	79.6%	3.9%	8.3%	8.3%	100.0%
	% within Classification	30.4%	100.0%	65.2%	41.7%	33.5%
Between 100 and 200 Cases	Count	236	0	7	16	259
	% within Group Size	91.1%	.0%	2.7%	6.2%	100.0%
	% within Classification	49.8%	.0%	30.4%	44.4%	48.0%
Over 200 Cases	Count	94	0	1	5	100
	% within Group Size	94.0%	.0%	1.0%	5.0%	100.0%
	% within Classification	19.8%	.0%	4.3%	13.9%	18.5%
Total Items	Count	474	7	23	36	540
	% within Group Size	87.8%	1.3%	4.3%	6.7%	100.0%
	% within Classification	100.0%	100.0%	100.0%	100.0%	100.0%

Focal Group Responses	Counts and Percents	Interaction Effect Size				Total
		No DIF	Large	Moderate	Small	
Below 100 Cases	Count	145	5	12	19	181
	% within Group Size	80.1%	2.8%	6.6%	10.5%	100.0%
	% within Classification	29.5%	100.0%	70.6%	70.4%	33.5%
Between 100 and 200 Cases	Count	247	0	5	7	259
	% within Group Size	95.4%	.0%	1.9%	2.7%	100.0%
	% within Classification	50.3%	.0%	29.4%	25.9%	48.0%
Over 200 Cases	Count	99	0	0	1	100
	% within Group Size	99.0%	.0%	.0%	1.0%	100.0%
	% within Classification	20.2%	.0%	.0%	3.7%	18.5%
Total Items	Count	491	5	17	27	540
	% within Group Size	90.9%	.9%	3.1%	5.0%	100.0%
	% within Classification	100.0%	100.0%	100.0%	100.0%	100.0%

Table 11
Grade 6 Items

Focal Group Responses	Counts and Percents	Uniform Effect Size				Total
		No DIF	Large	Moderate	Small	
Below 100 Cases	Count	130	14	21	28	193
	% within Group Size	67.4%	7.3%	10.9%	14.5%	100.0%
	% within Classification	40.0%	100.0%	77.8%	59.6%	46.7%
Between 100 and 200 Cases	Count	98	0	4	8	110
	% within Group Size	89.1%	.0%	3.6%	7.3%	100.0%
	% within Classification	30.2%	.0%	14.8%	17.0%	26.6%
Over 200 Cases	Count	97	0	2	11	110
	% within Group Size	88.2%	.0%	1.8%	10.0%	100.0%
	% within Classification	29.8%	.0%	7.4%	23.4%	26.6%
Total Items	Count	325	14	27	47	413
	% within Group Size	78.7%	3.4%	6.5%	11.4%	100.0%
	% within Classification	100.0%	100.0%	100.0%	100.0%	100.0%

Focal Group Responses	Counts and Percents	Interaction Effect Size				Total
		No DIF	Large	Moderate	Small	
Below 100 Cases	Count	132	18	22	21	193
	% within Group Size	68.4%	9.3%	11.4%	10.9%	100.0%
	% within Classification	38.3%	100.0%	88.0%	84.0%	46.7%
Between 100 and 200 Cases	Count	104	0	3	3	110
	% within Group Size	94.5%	.0%	2.7%	2.7%	100.0%
	% within Classification	30.1%	.0%	12.0%	12.0%	26.6%
Over 200 Cases	Count	109	0	0	1	110
	% within Group Size	99.1%	.0%	.0%	.9%	100.0%
	% within Classification	31.6%	.0%	.0%	4.0%	26.6%
Total Items	Count	345	18	25	25	413
	% within Group Size	83.5%	4.4%	6.1%	6.1%	100.0%
	% within Classification	100.0%	100.0%	100.0%	100.0%	100.0%

Table 12
Grade 7 Items

Focal Group Response	Counts and Percents	Uniform Effect Size				Total
		No DIF	Large	Moderate	Small	
Below 100 Cases	Count	133	16	14	22	185
	% within Group Size	71.9%	8.6%	7.6%	11.9%	100.0%
	% within Classification	35.8%	84.2%	66.7%	62.9%	41.4%
Between 100 and 200 Cases	Count	165	2	4	10	181
	% within Group Size	91.2%	1.1%	2.2%	5.5%	100.0%
	% within Classification	44.4%	10.5%	19.0%	28.6%	40.5%
Over 200 Cases	Count	74	1	3	3	81
	% within Group Size	91.4%	1.2%	3.7%	3.7%	100.0%
	% within Classification	19.9%	5.3%	14.3%	8.6%	18.1%
Total Items	Count	372	19	21	35	447
	% within Group Size	83.2%	4.3%	4.7%	7.8%	100.0%
	% within Classification	100.0%	100.0%	100.0%	100.0%	100.0%

Focal Group Response	Counts and Percents	Interaction Effect Size				Total
		No DIF	Large	Moderate	Small	
Below 100 Cases	Count	139	5	18	23	185
	% within Group Size	75.1%	2.7%	9.7%	12.4%	100.0%
	% within Classification	35.8%	100.0%	85.7%	69.7%	41.4%
Between 100 and 200 Cases	Count	168	0	3	10	181
	% within Group Size	92.8%	.0%	1.7%	5.5%	100.0%
	% within Classification	43.3%	.0%	14.3%	30.3%	40.5%
Over 200 Cases	Count	81	0	0	0	81
	% within Group Size	100.0%	.0%	.0%	.0%	100.0%
	% within Classification	20.9%	.0%	.0%	.0%	18.1%
Total Items	Count	388	5	21	33	447
	% within Group Size	86.8%	1.1%	4.7%	7.4%	100.0%
	% within Classification	100.0%	100.0%	100.0%	100.0%	100.0%

Table 13
Grade 8 Items

Focal Group Response	Counts and Percents	Uniform Effect Size				Total
		No DIF	Large	Moderate	Small	
Below 100 Cases	Count	352	80	63	53	548
	% within Group Size	64.2%	14.6%	11.5%	9.7%	100.0%
	% within Classification	73.5%	98.8%	96.9%	77.9%	79.1%
Between 100 and 200 Cases	Count	119	1	2	15	137
	% within Group Size	86.9%	.7%	1.5%	10.9%	100.0%
	% within Classification	24.8%	1.2%	3.1%	22.1%	19.8%
Over 200 Cases	Count	8	0	0	0	8
	% within Group Size	100.0%	.0%	.0%	.0%	100.0%
	% within Classification	1.7%	.0%	.0%	.0%	1.2%
Total Items	Count	479	81	65	68	693
	% within Group Size	69.1%	11.7%	9.4%	9.8%	100.0%
	% within Classification	100.0%	100.0%	100.0%	100.0%	100.0%

Focal Group Response	Counts and Percents	Interaction Effect Size				Total
		No DIF	Large	Moderate	Small	
Below 100 Cases	Count	366	69	55	58	548
	% within Group Size	66.8%	12.6%	10.0%	10.6%	100.0%
	% within Classification	73.3%	100.0%	96.5%	85.3%	79.1%
Between 100 and 200 Cases	Count	125	0	2	10	137
	% within Group Size	91.2%	.0%	1.5%	7.3%	100.0%
	% within Classification	25.1%	.0%	3.5%	14.7%	19.8%
Over 200 Cases	Count	8	0	0	0	8
	% within Group Size	100.0%	.0%	.0%	.0%	100.0%
	% within Classification	1.6%	.0%	.0%	.0%	1.2%
Total Items	Count	499	69	57	68	693
	% within Group Size	72.0%	10.0%	8.2%	9.8%	100.0%
	% within Classification	100.0%	100.0%	100.0%	100.0%	100.0%

Table 14
Grade 10 Items

Jodoin/Gierl Effect Size	ETS DIF Classification			Total
	A	B	C	
No DIF	458	32	3	493
Moderate	0	0	8	8
Small	5	10	15	30
Total	463	42	26	531

Table 15
Grade 3 Comparison
Over 100 Students in Each Group

Jodoin/Gierl Effect Size	ETS DIF Classification			Total
	A	B	C	
No DIF	377	31	4	412
Large	0	0	1	1
Moderate	0	0	8	8
Small	3	9	19	31
Total	380	40	32	452

Table 16
Grade 4 Comparison
Over 100 Students in Each Group

Jodoin/Gierl Effect Size	ETS DIF Classification			Total
	A	B	C	
No DIF	368	28	1	397
Large	0	0	1	1
Moderate	0	1	7	8
Small	4	8	16	28
Total	372	37	25	434

Table 17
Grade 5 Comparison
Over 100 Students in Each Group

Jodoin/Gierl Effect Size	ETS DIF Classification			
	A	B	C	Total
No DIF	316	10	4	330
Moderate	0	0	8	8
Small	8	6	7	21
Total	324	16	19	359

Table 18
Grade 6 Comparison
Over 100 Students in Each Group

Jodoin/Gierl Effect Size	ETS DIF Classification			
	A	B	C	Total
No DIF	181	14	0	195
Moderate	0	0	6	6
Small	2	6	11	19
Total	183	20	17	220

Table 19
Grade 7 Comparison
Over 100 Students in Each Group

Jodoin/Gierl Effect Size	ETS DIF Classification			
	A	B	C	Total
No DIF	224	14	1	239
Large	0	0	3	3
Moderate	1	0	6	7
Small	5	3	5	13
Total	230	17	15	262

Table 20
Grade 8 Comparison
Over 100 Students in Each Group

Jodoin/Gierl Effect Size	ETS DIF Classification			
	A	B	C	Total
No DIF	124	2	1	127
Large	0	0	1	1
Moderate	0	0	2	2
Small	5	2	8	15
Total	129	4	12	145

Table 21

Grade 10 Comparison

Over 100 Students in Each Group

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington D.C.: American Psychological Association.

Camilli, G. (2006). Test fairness. In Brennan, R.L. *Educational Measurement*, (4th ed.), Westport, CT: Praeger Publishers.

Crocker and Algina (1986). *Introduction to classical and modern test theory*. New York: Harcourt Brace Javanovich College Publishers.

Fidalgo, Ferreres, & Muniz. (2004). Utility of the mantel-haenszel procedure for detecting differential item functioning in small samples. *Educational and Psychological Measurement, 64*, 925-936.

Holland, P. W. (1985). On the study of differential item performance without IRT. Proceedings of the Military Testing Association, October.

Holland, P. W. & Thayer, D.T. (1985). An alternative definition of the ETS delta scale of item difficulty. Princeton, N.J.: Educatational Testing Service, Research Report, RR-85-43.

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

Jodoin, M.G., & Gierl, M.J. (2001). Evaluating power and type I error rates using an effect size with the logistic regression procedure for DIF. *Applied Measurement in Education*, 14, 329-349.

Koretz, D.M. and Hamilton, L.S (2006). Testing for accountability in K-12. (4th ed., pp. 531-578). In Brennan, R.L. Educational Measurement. (Westport, CT: Greenwood Publishers).

Mazor, K.M., Clauser, B.E. & Hambleton, R.K. (1992). The effect of sample size on thefunctioning of the Mantel-Haenszel statistic. Educational and Psychological Measurement, 52, 443-451.

Stienberg, L., Thissen, D., & Wainer, H. (1990). Validity. In H. Wainer (Ed.), Computerized adaptive testing: A primer (pp.187-231. Hillsdale, N.J.: Erlbaum.

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

Zwick , R. (2000). The assessment of differential item functioning in computer-adaptive tests. Chapter in *Computerized Adaptive Testing: Theory and Practice*, W. J. van der Linden & C. A. W. Glas, Eds., Kluwer, 2000.