

To: Michael J. Muenks, Shaun Bates

**Dept: Missouri Department of
Elementary and Secondary Education**

Location: Missouri

**Subject: Missouri Assessment Program
Preliminary Differential Item Functioning
(DIF) Report (Version 2)**

From DRC: Joanna Tomkowicz

Dept: Psychometric Services

Location: Maple Grove, Minnesota

Date: July 30, 2015

Missouri Assessment Program Preliminary Differential Item Functioning (DIF) Report Version 2

This document outlines procedure implemented to conduct differential item (DIF) analysis on 2014-2015 Missouri Assessment Program (MAP) data. The DIF statistics indicate the degree to which members of a particular subgroup performs better or worse than expected on each item as compared to the reference group.

Version History

This document contains data tables that were previously delivered to DESE separately (in an excel file) and replaces the original document with the same title dated July 10, 2015. No other changes were made.

Background and methodology

Our position concerning test bias is based on two general propositions. First, students may differ in their background knowledge, cognitive and academic skills, language, attitudes, and values. To the degree that these differences are large, no one curriculum and no one set of instructional materials will be equally suitable for all. Therefore, no one test will be equally appropriate for all. Furthermore, it is difficult to specify what amount of difference can be called large and to determine how these differences will affect the outcome of a particular test. Second, schools have been assigned the tasks of developing certain basic cognitive skills and supporting development of these skills equitably among all students. Therefore, there is a need for tests that measure the common skills and bodies of knowledge that are common to all learners. The test publisher's task is to develop assessments that measure these key cognitive skills without introducing extraneous or construct-irrelevant elements into the performances on which the measurement is based. If these tests require that students have culturally-specific knowledge and skills not taught in school, differences in performance among students can occur because of differences in student background and out-of-school learning. Such tests are measuring different things for different groups and can be called biased (Camilli & Shepard, 1994; Green, 1975).

In order to lessen this bias, we strive to minimize the role of extraneous elements, thereby increasing the number of students for whom the test is appropriate. Therefore, careful attention is

given during the test development and test construction processes to lessen the influence of these elements for large numbers of students (including the use of Bias Review committees). Unfortunately, in some cases these elements may continue to play a substantial role. To assess the extent to which items may be performing differently for various subgroups of interest, DIF analyses are conducted after each operational test administration.

DIF statistics are used to quantify differences in item performance between two groups after controlling for examinees' overall achievement level. Two DIF statistics that are commonly used for this purpose are the Mantel-Haenszel (MH) statistic (1959) and the Standardized Mean Difference (SMD) between the reference and focal groups, proposed by Dorans and Schmitt (1991).

The MH statistic is computed as (Zwick, Donoghue, & Grima, 1993):

$$\text{Mantel } \chi^2 = \frac{\left(\sum_k F_k - \sum_k E(F_k) \right)^2}{\sum_k \text{Var}(F_k)},$$

where F_k is the sum of scores for the focal group at the k^{th} level of the matching variable. Note that the MH statistic is sensitive to N such that larger sample sizes increase the value of chi square.

In addition to the MH chi-square statistic, the delta statistic (MH-D DIF) was computed for all items. Educational Testing Service (ETS) first developed the MH-D DIF statistic. To compute delta, alpha (the odds ratio) is first computed as:

$$\alpha_{MH} = \frac{\sum_{k=1}^K N_{r1k}N_{f0k} / N_k}{\sum_{k=1}^K N_{f1k}N_{r0k} / N_k},$$

where N_{r1k} is the number of correct responses in the reference group at ability level k , N_{f0k} is the number of incorrect responses in the focal group at ability level k , N_k is the total number of responses, N_{f1k} is the number of correct responses in the focal group at ability level k , and N_{r0k} is the number of incorrect responses in the reference group at ability level k . MH-D DIF is then computed as:

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

For selected-response items, the MH (χ_{MH}^2) statistic was used to evaluate potential DIF items. In the MH procedure, subgroups are matched by their raw total test score, using a contingency table with K ability levels. When applying the MH procedure, the log-odds ratio α is assumed to be constant across the K matched levels. The χ_{MH}^2 , then, estimates a pooled common-odds ratio. Taking the natural logarithm of the common-odds ratio and its confidence limits and multiplying

these with the constant -2.35 , the resulting values may then be placed on the MH delta metric (Δ_{MH}) for interpretive purposes. Items were flagged for DIF using the following criteria:

- Moderate DIF: Significant MH chi-square statistic ($p < 0.05$) and $1.0 \leq |\text{MH D-DIF}| < 1.5$
- Large DIF: Significant MH chi-square statistic ($p < 0.05$) and $|\text{MH D-DIF}| \geq 1.5$

For constructed-response items, an effect size (ES) statistic based on the MH chi-square will be used. The ES is obtained by dividing the SMD statistics by the standard deviation of the item. The SMD is an effect size index of DIF, which is relatively easy to interpret (Zwick et al., 1993). The SMD compares the mean of the reference and focal group, adjusting for the distribution of reference and focal group members on the conditioning variable (Zwick et al., 1993), which for these analyses is the MAP raw score. SMD is computed as (Zwick et al., 1993):

$$SMD = p_{Fk} \left(\sum_k m_{Fk} - \sum_k m_{Rk} \right),$$

where p_{Fk} = proportion of the focal group members at the k th level of the matching variable, $m_{Fk} = 1/N_{F1k}$, and $m_{Rk} = 1/N_{R1k}$. Items are flagged using the same rules that are used in NAEP:

- Moderate DIF: If the MH statistic is significant ($p < .05$) and $|\text{ES}|$ is between 0.17 and 0.25.
- Large DIF: If the MH statistic is significant ($p < .05$) and $|\text{ES}| \geq 0.25$.

A positive DIF value indicates that the item favors the focal group, while a negative value indicates that the item disadvantages the focal group. In this document DIF results are presented for the following subgroups:

- **Gender:** Focal group is Females; Reference group is Males.
- **Ethnicity:** Focal groups are Black (African-American), Hispanic, Asian/Pacific Islander, and Other; Reference group is White.
- **Accommodations:** Focal group is students who received one or more testing accommodations in a given content area; Reference group is all others.

A negative SMD value implies that the focal group has a lower mean item score than the reference group, whereas a positive value implies that the focal group has a higher mean item score than the reference group, conditioned on the matching test score.

The DIF analyses on the 2014-2015 MAP data were performed by test form within a grade and content area. The core forms (CA) and performance assessment forms (PA) for grades 5 and 8 ELA and Mathematics were analyzed separately.

The minimum case count for the focal group was set at 200 and the minimum case count for the reference group was set at 400. These criteria were used in MAP DIF analyses in previous administrations.

Results

The tables in this section summarize the number of DIF flags by test form, grade, and content area for each focal group versus reference group. The tables also show the number of items on each test. For example, let's consider Grade 3 ELA, form CA1 (see Table 1). In this form, four items were flagged for DIF for the Asian/Pacific Islander subgroup. Of these items, three items exhibited moderate negative DIF and one item exhibited large positive DIF. Two items were flagged for the African-American subgroup: one displaying moderate negative DIF and one showing moderate positive DIF. Two items were flagged for Hispanic subgroup: one displaying moderate negative DIF and one showing large negative DIF. Note that DIF analysis was not performed for the American Indian/Alaska Native subgroup due to fewer than 200 students in a focal group. Three items were flagged for DIF for the female subgroup: all three exhibited moderate positive DIF. Lastly, two items were flagged for the accommodated subgroup: one of them exhibited moderate negative DIF and the other one showed moderate positive DIF.

It should be noted that in 2014-15 multiple test forms were administered within one grade and content area. Consequently, the number of American Indian/ Alaska Native students taking each form was smaller than 200 and no DIF was performed for this group on any test form. In addition, fewer than 200 Asian/ Pacific Islander students took form CA3 in Grade 8 Mathematics and no DIF was performed for this group using the data for this test form.

Core (CA) and performance assessment (PA) forms were analyzed separately in order to obtain at least 200 cases for minority subgroups. Conducting DIF analyses on the combined core and performance assessment forms would result in more groups excluded from DIF analyses.

Table 1. Number of Items Flagged for DIF: ELA Grades 3 and 4

ELA		Grade	3			4		
		Form	CA1	CA2	CA3	CA1	CA2	CA3
		No. items	44	44	44	44	44	44
Focal Group	DIF Magnitude	DIF Direction	Number of flagged items					
Asian/ Pacific Islander	Moderate	Negative	3	2	2	0	3	1
		Positive	0	5	4	0	1	1
	Large	Negative	0	1	0	0	0	1
		Positive	1	0	0	0	0	1
Black	Moderate	Negative	1	0	0	0	0	0
		Positive	1	0	0	0	0	1
	Large	Negative	0	0	0	0	0	0
		Positive	0	0	0	0	0	1
Hispanic	Moderate	Negative	1	1	1	0	0	0
		Positive	0	0	0	0	0	1
	Large	Negative	1	0	0	0	0	0
		Positive	0	0	0	0	0	0
Other	Moderate	Negative	0	0	0	0	0	0
		Positive	0	0	0	0	0	0
	Large	Negative	0	0	0	0	0	0
		Positive	0	0	0	0	0	0
Female	Moderate	Negative	0	0	0	0	2	0
		Positive	3	2	1	2	1	4
	Large	Negative	0	0	0	0	0	0
		Positive	0	0	0	1	0	2
Accommodated	Moderate	Negative	1	0	0	1	1	0
		Positive	1	0	0	0	0	0
	Large	Negative	0	0	0	0	0	0
		Positive	0	0	0	0	0	0

Table 2. Number of Items Flagged for DIF: ELA Grade 5

ELA		Grade	5					
		Form	CA1	CA2	CA3	PA1	PA2	PA3
		No. items	44	44	44	4	4	4
Focal Group	DIF Magnitude	DIF Direction	Number of flagged items					
Asian/Pacific Islander	Moderate	Negative	1	1	2	0	0	0
		Positive	2	0	1	0	0	0
	Large	Negative	1	0	0	0	0	0
		Positive	0	0	0	0	0	0
Black	Moderate	Negative	0	2	1	0	0	0
		Positive	1	0	0	0	0	0
	Large	Negative	0	0	0	0	0	0
		Positive	0	0	0	0	0	0
Hispanic	Moderate	Negative	1	1	1	0	0	0
		Positive	0	0	0	0	0	0
	Large	Negative	0	0	0	0	0	0
		Positive	0	0	0	0	0	0
Other	Moderate	Negative	0	0	1	0	0	0
		Positive	0	0	0	0	0	0
	Large	Negative	0	0	0	0	0	0
		Positive	0	0	0	0	0	0
Female	Moderate	Negative	1	3	3	0	0	0
		Positive	2	0	1	0	0	0
	Large	Negative	0	0	0	0	0	0
		Positive	1	2	1	0	0	0
Accommodated	Moderate	Negative	0	0	0	0	0	0
		Positive	0	0	0	0	0	0
	Large	Negative	0	0	0	0	0	0
		Positive	0	0	0	0	0	0

Table 3. Number of Items Flagged for DIF: ELA Grades 6 and 7

ELA		Grade	6			7		
		Form	CA1	CA2	CA3	CA1	CA2	CA3
		No. items	45	45	45	45	45	45
Focal Group	DIF	DIF	Number of flagged items					
Asian/Pacific Islander	Moderate	Negative	1	3	1	1	0	2
		Positive	3	0	1	1	2	1
	Large	Negative	0	0	1	0	1	0
		Positive	0	0	0	0	1	1
Black	Moderate	Negative	1	0	2	0	1	0
		Positive	0	0	1	0	1	0
	Large	Negative	0	0	0	0	0	0
		Positive	0	0	0	0	0	0
Hispanic	Moderate	Negative	1	0	1	0	2	0
		Positive	0	0	0	0	1	0
	Large	Negative	0	0	0	0	0	0
		Positive	0	0	0	0	0	0
Other	Moderate	Negative	0	0	0	0	0	0
		Positive	0	0	0	0	0	0
	Large	Negative	0	0	0	0	0	0
		Positive	0	0	0	0	0	0
Female	Moderate	Negative	2	1	0	1	1	3
		Positive	0	2	2	2	2	1
	Large	Negative	0	0	0	0	0	0
		Positive	2	1	1	1	1	2
Accommodated	Moderate	Negative	0	0	0	0	0	0
		Positive	0	0	0	0	0	0
	Large	Negative	0	0	0	0	0	0
		Positive	0	0	0	0	0	0

Table 4. Number of Items Flagged for DIF: ELA Grade 8

ELA		Grade	8					
		Form	CA1	CA2	CA3	PA1	PA2	PA3
		No. items	45	45	45	4	4	4
Focal Group	DIF Magnitude	DIF Direction	Number of items flagged		Number of flagged items			
Asian/Pacific Islander	Moderate	Negative	4	2	2	0	0	0
		Positive	1	2	2	0	0	0
	Large	Negative	0	0	0	0	0	0
		Positive	1	0	0	0	0	0
Black	Moderate	Negative	0	0	0	0	0	0
		Positive	1	0	0	0	0	0
	Large	Negative	0	0	0	0	0	0
		Positive	0	0	0	0	0	0
Hispanic	Moderate	Negative	1	0	0	0	0	0
		Positive	0	0	0	0	0	0
	Large	Negative	0	0	0	0	0	0
		Positive	0	0	0	0	0	0
Other	Moderate	Negative	0	0	0	0	0	0
		Positive	0	0	0	0	0	0
	Large	Negative	0	0	0	0	0	0
		Positive	0	0	0	0	0	0
Female	Moderate	Negative	3	5	2	0	0	0
		Positive	1	0	1	0	0	0
	Large	Negative	0	1	0	0	0	0
		Positive	3	3	2	0	0	0
Accommodated	Moderate	Negative	0	1	0	0	0	0
		Positive	0	0	0	0	0	0
	Large	Negative	0	0	0	0	0	0
		Positive	0	0	0	0	0	0

Table 5. Number of Items Flagged for DIF: Mathematics Grades 3 and 4

Mathematics		Grade	3			4		
		Form	CA1	CA2	CA3	CA1	CA2	CA3
		No. items	31	31	31	31	31	31
Focal Group	DIF	DIF	Number of flagged items					
Asian/Pacific Islander	Moderate	Negative	0	0	0	0	0	0
		Positive	0	0	0	0	0	1
	Large	Negative	0	0	0	0	0	0
		Positive	0	0	0	0	0	0
Black	Moderate	Negative	1	1	0	1	2	2
		Positive	0	0	0	0	2	3
	Large	Negative	0	1	0	0	0	0
		Positive	0	0	0	0	0	0
Hispanic	Moderate	Negative	1	1	0	0	0	0
		Positive	0	0	0	0	0	0
	Large	Negative	0	0	0	0	0	0
		Positive	0	0	0	0	0	0
Other	Moderate	Negative	0	0	0	0	0	0
		Positive	0	0	0	0	0	0
	Large	Negative	0	0	0	0	0	0
		Positive	0	0	0	0	0	0
Female	Moderate	Negative	0	1	0	0	0	0
		Positive	0	0	0	0	0	0
	Large	Negative	0	0	0	0	0	0
		Positive	0	0	0	0	0	0
Accommodated	Moderate	Negative	0	0	0	0	0	0
		Positive	0	0	0	0	0	0
	Large	Negative	0	0	0	0	0	0
		Positive	0	0	0	0	0	0

Table 6. Number of Items Flagged for DIF: Mathematics Grades 5 and 6

Mathematics		Grade	5	5	5	5	5	6	6	6	
		Form	CA1	CA2	CA3	PA1	PA2	CA1	CA2	CA3	
		No. items	31	31	31	6	6	30	30	30	
Focal Group	DIF	DIF	Number of flagged items								
Asian/Pacific Islander	Moderate	Negative	0	1	0	0	0	0	0	1	1
		Positive	0	0	0	0	0	0	0	1	2
	Large	Negative	0	0	0	0	0	1	0	0	0
		Positive	0	0	0	0	0	0	0	0	0
Black	Moderate	Negative	1	1	1	0	0	0	0	0	0
		Positive	1	0	0	0	0	0	0	1	2
	Large	Negative	0	0	1	0	0	0	0	0	0
		Positive	0	0	0	0	0	0	0	0	0
Hispanic	Moderate	Negative	0	0	0	0	0	0	0	0	0
		Positive	0	0	0	0	0	0	0	0	0
	Large	Negative	0	0	0	0	0	0	0	0	0
		Positive	0	0	0	0	0	0	0	0	0
Other	Moderate	Negative	0	0	0	0	0	0	0	0	0
		Positive	0	0	0	0	0	0	0	0	0
	Large	Negative	0	0	0	0	0	0	0	0	0
		Positive	0	0	0	0	0	0	0	0	0
Female	Moderate	Negative	0	0	0	0	0	1	0	1	
		Positive	0	0	0	0	0	0	0	0	
	Large	Negative	0	0	0	0	0	0	0	0	
		Positive	0	0	0	0	0	0	0	0	
Accommodated	Moderate	Negative	0	0	0	0	0	0	0	0	
		Positive	0	0	0	0	0	1	0	0	
	Large	Negative	0	0	0	0	0	0	0	0	
		Positive	0	0	0	0	0	0	0	0	

Table 7. Number of Items Flagged for DIF: Mathematics Grades 7 and 8

Mathematics		Grade	7	7	7	8	8	8	8	8
		Form	CA1	CA2	CA3	CA1	CA2	CA3	PA1	PA2
		No. items	31	31	31	31	31	31	6	6
Focal Group	DIF	DIF	Number of flagged items							
Asian/Pacific Islander	Moderate	Negative	1	1	0	0	1	n/a	0	0
		Positive	0	1	0	0	1	n/a	0	0
	Large	Negative	1	0	0	0	0	n/a	0	0
		Positive	0	0	0	0	0	n/a	0	0
Black	Moderate	Negative	0	1	1	0	0	1	0	0
		Positive	1	0	0	1	0	1	0	0
	Large	Negative	0	0	1	0	0	0	0	0
		Positive	0	0	0	0	0	0	0	0
Hispanic	Moderate	Negative	0	0	0	0	0	0	0	0
		Positive	0	0	0	0	0	0	0	0
	Large	Negative	0	0	0	0	0	0	0	0
		Positive	0	0	0	0	0	0	0	0
Other	Moderate	Negative	0	0	0	0	0	0	0	0
		Positive	0	0	0	0	0	0	0	0
	Large	Negative	0	0	0	0	0	0	0	0
		Positive	0	0	0	0	0	0	0	0
Female	Moderate	Negative	0	1	1	0	0	0	0	0
		Positive	1	0	0	0	0	0	0	0
	Large	Negative	0	0	0	0	0	0	0	0
		Positive	0	0	0	0	0	0	0	0
Accommodated	Moderate	Negative	0	0	0	0	0	0	0	0
		Positive	1	2	2	0	1	0	0	0
	Large	Negative	0	0	0	0	0	0	0	0
		Positive	0	0	0	0	0	0	0	0

Table 8. Number of Items Flagged for DIF: Science Grades 5 and 8

Science		Grade	5	5	8	8
		Form	CA2	CA3	CA2	CA3
		No. items	41	41	39	38
Focal Group	DIF Magnitude	DIF Direction	Number of flagged items			
Asian/Pacific Islander	Moderate	Negative	3	2	1	1
		Positive	0	1	1	0
	Large	Negative	0	0	0	0
		Positive	0	0	0	1
Black	Moderate	Negative	0	4	1	0
		Positive	0	1	1	2
	Large	Negative	0	0	0	0
		Positive	0	0	0	0
Hispanic	Moderate	Negative	1	1	0	0
		Positive	0	1	0	0
	Large	Negative	0	0	0	0
		Positive	0	0	0	0
Other	Moderate	Negative	0	0	0	0
		Positive	0	0	0	0
	Large	Negative	0	0	0	0
		Positive	0	0	0	0
Female	Moderate	Negative	1	1	3	1
		Positive	0	0	2	1
	Large	Negative	0	0	0	0
		Positive	0	0	1	1
Accommodated	Moderate	Negative	0	0	0	1
		Positive	1	0	0	0
	Large	Negative	0	0	0	0
		Positive	0	0	0	0

Summary and recommendation

In order to address these limitations we propose the following potential solutions for DESE consideration:

- 1) Setting minimum case count for a focal group at 100. This option will allow for conducting DIF analysis for the Asian/Pacific Islander group on form CA3 in Grade 8 Mathematics and on a few test forms for the American Indian/Alaska Native subgroup. However, the number of American Indian/Alaska Native student taking most of the test forms will still be less than 100. The DIF analyses should not be performed for subgroups of less than 100. In such cases, the statistical procedures do not have sufficient power to detect differences should they exist.
- 2) Conducting DIF analyses on a subset of items for the American Indian/Alaska Native subgroup. In this option, DIF analysis would be conducted on a subset of items common across core forms within a grade and a content area. These item collections would be treated as complete test forms and the total test scores on these item collections would be used to control for students' overall achievement. Limiting the number of items included in DIF analysis to common items across forms would allow for conducting DIF analysis on these items for the American Indian/ Alaska Native students. DIF analysis for this subgroup would not be conducted on any other items for which the case count is less than 100.

In summary, it should be noted that any items included on the MAP (including those items flagged for DIF) have been thoroughly reviewed for content and bias by Missouri educators and CTB Content Development staff. The DIF flags do not necessarily indicate that an item is biased; rather, DIF flags indicate that the item functions differently for equally able members of different groups (Camilli & Shepard, 1994). All items flagged for DIF shown in the excel file had been reviewed before inclusion on the operational MAP to insure that they do not tap knowledge or specific ability irrelevant to the construct the test intends to measure. Items are not necessarily suppressed from operational scoring if they are flagged for DIF.

References

- Camilli, G., & Shepard, A. L. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA :Sage Publication.
- Dorans, N.J., & Schmitt, M.P. (1991). *Constructed response and differential item functioning: A pragmatic approach*. Princeton: Educational Testing Service.
- Green, D.R. (1975). Procedures for assessing bias in achievement tests. Paper presented at the National Institute of Education Conference on Test Bias, Annapolis, MD.
- Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22, 719–748.
- Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement*, 30, 233–251.