Spring 2016 MAP
Science
Early Tables

DRAFT

DRAFT

Developed and published under contract with Missouri Department of Elementary and Secondary Education (DESE) by Data Recognition Corporation, 13490 Bass Lake Road, Maple Grove, MN 55311. Copyright © 2016 by Missouri DESE. All rights reserved. Only Missouri State educators and citizens may copy and/or download and print the document, located online at http://dese.mo.gov/college-career-readiness/assessment/assessment-technical-support-materials. Any other use or reproduction of this document, in whole or in part, requires the prior written permission of Missouri DESE.

Table of Contents

CHAPTER 5: SCORING OF CONSTRUCTED-RESPONSE ITEMS 4
5.1. Inter-Rater Reliability 4
CHAPTER 7: TEST RESULTS 7
7.1. Cross-year, Cross-sectional Comparisons 7
CHAPTER 10: FAIRNESS 10
10.1. Evaluating Bias through Differential Item Functioning (DIF) Statistics 10
10.2. Evaluating Bias through Impact Analysis 13
10.2.1. Effect Size 13
10.2.2. Reliability 14
References 19

List of Tables

Table 5.1: Inter-rater Reliability, Science Grade 5 5
Table 5.2: Inter-rater Reliability, Science Grade 8 6
Table 7.1. Comparison of State-Level Means, 2008 through 2016: Science 8
Table 7.2. Comparison of Percentage of Students in each Achievement Level, Science 2008 through 2016 Census Data 9
Table 10.1: 2015 MAP DIF Statistics: Number of Flagged Items, Science 15
Table 10.2: Impact Analysis, Grade 5 Science 16
Table 10.3: Impact Analysis, Grade 8 Science 16
Table 10.4: Grade 5 Science Reliability and SEM by Subgroup 17
Table 10.5: Grade 8 Science Reliability and SEM by Subgroup 18

DRAFT

This document contains early results of the 2016 Science Grades 5 and 8 summative assessments administered to Missouri students in Spring 2016. Early results include: inter-rater reliability statistics (Chapter 5); state-level impact data and longitudinal impact data (Chapter 7); differential item functioning analysis, scale score summaries by subgroups (gender, ethnicity, and using or not using testing accommodations), and test reliability coefficients and standard error of measurement for total population and by subgroups (Chapter 10). The chapter numbers in this document are not consecutive and reflect the chapter numbers in the subsequent Spring 2016 Science Technical Report.

CHAPTER 5: SCORING OF CONSTRUCTED-RESPONSE ITEMS

5.1. Inter-Rater Reliability

Approximately 10% of the papers in Science were scored independently by a second reader. The statistics for the inter-rater reliability were calculated for all items at all grades. To determine the reliability of scoring, the percentage of perfect agreement and adjacent agreement between the two readers was examined.

For each item, a quadratic weighted kappa was calculated to reflect the level of improvement beyond the chance level in the consistency of scoring. These quadratic weighted kappa values are presented in Table 5.1 and Table 5.2 for Grades 5 and 8, respectively. To aid in the interpretation of kappa, the following cutoffs have been suggested (Landis & Koch, 1977; Altman, 1991):

Kappa Value	Strength of Agreement
0	None
<0.20	Poor
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Good
0.81 – 1.00	Very Good

A total of 69 items were scored by human readers across all test forms for both Science grade levels. As shown in Table 5.1 and 5.2, raters demonstrated above 97% perfect and adjacent agreement for all Science items. The quadratic weighted kappa ranged from 0.60 to 0.98 indicting that there was good or very good inter-rater agreement for all Science items.

DRAFT

Table 5.1: Inter-rater Reliability, Science Grade 5

Grade	Session	Form	Item #	# Points	% Perfect	% Adjacent	% Perfect & Adjacent*	Quadratic Weighted Kappa
	1	CA2	1	2	88.2%	11.3%	99.6%	0.89
	1	CA2	3	2	85.9%	13.2%	99.1%	0.83
	1	CA2	4	2	87.9%	11.8%	99.7%	0.89
	1	CA2	7	2	93.1%	6.8%	99.9%	0.92
	1	CA2	8	2	72.4%	25.6%	98.0%	0.70
	1	CA2	10	2	92.4%	7.6%	100.0%	0.93
	1	CA2	11	2	93.5%	6.3%	99.8%	0.94
	1	CA2	12	2	83.0%	17.0%	100.0%	0.82
	1	CA2	13	2	96.8%	3.2%	100.0%	0.96
	1	CA2	14	2	95.3%	4.5%	99.8%	0.95
	1	CA2, CA3	2	2	93.8%	6.1%	99.9%	0.94
	1	CA2, CA3	5	2	96.4%	3.3%	99.7%	0.96
	1	CA2, CA3	6	2	92.9%	7.0%	99.9%	0.93
	1	CA2, CA3	9	2	87.2%	12.6%	99.8%	0.85
	1	CA3	1	2	94.7%	5.3%	100.0%	0.95
	1	CA3	3	2	90.3%	9.6%	99.9%	0.85
5	1	CA3	4	2	98.1%	1.4%	99.4%	0.98
	1	CA3	7	2	99.0%	1.0%	100.0%	0.98
	1	CA3	8	2	92.2%	7.7%	100.0%	0.91
	1	CA3	10	2	88.9%	10.9%	99.9%	0.85
	1	CA3	11	2	89.9%	9.8%	99.8%	0.89
	1	CA3	12	2	95.9%	4.1%	100.0%	0.96
	1	CA3	13	2	88.9%	11.0%	99.9%	0.87
	1	CA3	14	2	95.8%	4.2%	100.0%	0.96
	3	CA2, CA3	33	2	99.1%	0.4%	99.6%	0.96
	3	CA2, CA3	34	4	90.5%	6.9%	97.5%	0.96
	3	CA2, CA3	35	1	93.2%	6.8%	100.0%	0.86
	3	CA2, CA3	36	1	97.9%	2.2%	100.0%	0.96
	3	CA2, CA3	37	2	75.7%	22.2%	98.0%	0.72
	3	CA2, CA3	38	1	89.7%	10.3%	100.0%	0.72
	3	CA2, CA3	39	1	88.7%	11.3%	100.0%	0.63
	3	CA2, CA3	40	1	89.6%	10.4%	100.0%	0.60
	3	CA2, CA3	41	1	98.4%	1.6%	100.0%	0.97

* The percent perfect & adjacent may not add up to 100 due to the percent discrepant (the cases where the assigned score varied by more than 1 point).

DRAFT

Table 5.2: Inter-rater Reliability, Science Grade 8

Grade	Session	Form	Item #	# Points	% Perfect	% Adjacent	% Perfect & Adjacent*	Quadratic Weighted Kappa
	1	CA2, CA3	5	2	95.9%	4.0%	99.6%	0.97
	1	CA2, CA3	10	2	96.9%	3.1%	99.1%	0.97
	1	CA2	1	2	83.2%	14.8%	99.7%	0.82
	1	CA2	2	2	82.4%	16.1%	99.9%	0.84
	1	CA2	3	2	92.3%	7.6%	98.0%	0.93
	1	CA2	4	2	97.6%	2.3%	100.0%	0.97
	1	CA2	6	2	88.2%	10.2%	99.8%	0.88
	1	CA2	7	2	83.8%	13.9%	100.0%	0.86
	1	CA2	8	2	84.6%	14.9%	100.0%	0.85
	1	CA2	9	2	90.8%	9.0%	99.8%	0.93
	1	CA2	11	2	92.9%	7.0%	99.9%	0.94
	1	CA2	12	2	91.3%	8.6%	99.7%	0.89
	1	CA2	13	2	90.8%	8.8%	99.9%	0.90
	1	CA2	14	2	92.2%	7.8%	99.8%	0.91
	1	CA3	1	2	84.1%	15.0%	100.0%	0.82
	1	CA3	2	2	80.7%	18.6%	99.9%	0.82
	1	CA3	3	2	77.1%	21.6%	99.4%	0.73
8	1	CA3	4	2	90.9%	8.8%	100.0%	0.92
	1	CA3	5	2	96.9%	2.8%	100.0%	0.97
	1	CA3	6	2	80.1%	18.0%	99.9%	0.83
	1	CA3	7	2	85.7%	14.0%	99.8%	0.86
	1	CA3	8	2	96.5%	3.5%	100.0%	0.72
	1	CA3	9	2	70.4%	26.1%	99.9%	0.69
	1	CA3	11	2	79.5%	19.4%	100.0%	0.81
	1	CA3	12	2	78.3%	20.4%	99.6%	0.77
	1	CA3	13	2	76.7%	21.9%	97.5%	0.74
	1	CA3	14	2	82.9%	16.4%	100.0%	0.81
	3	CA2, CA3	31, 30	2	94.9%	4.9%	100.0%	0.95
	3	CA2, CA3	32, 31	2	96.8%	3.0%	98.0%	0.98
	3	CA2, CA3	33, 32	1	87.4%	12.6%	100.0%	0.71
	3	CA2, CA3	34, 33	1	99.4%	0.6%	100.0%	0.98
	3	CA2, CA3	35, 34	4	79.8%	18.4%	100.0%	0.93
	3	CA2, CA3	36, 35	1	90.2%	9.8%	100.0%	0.80
	3	CA2, CA3	37, 36	2	86.4%	12.0%	100.0%	0.86
	3	CA2, CA3	38, 37	3	80.7%	16.3%	100.0%	0.88
	3	CA2, CA3	39, 38	1	88.5%	11.6%	98.0%	0.75

* The percent perfect & adjacent may not add up to 100 due to the percent discrepant (the cases where the assigned score varied by more than 1 point).

DRAFT

CHAPTER 7: TEST RESULTS

This chapter of the Technical Report contains information on the results of the Spring 2016 administration of the Science MAP. The scale score results are presented here. Achievement-level information is also provided. Presenting the results by achievement level translates the quantitative scale provided through scale scores into a qualitative description of student achievement: *Below Basic*, *Basic*, *Proficient*, and *Advanced.*

While the scale score provides an essential quantitative reference to student achievement, the achievement-level information plainly outlines the meaning of the scores to parents, students, and educators. When combined, scale scores and achievement levels provide a comprehensive set of tools to assess Missouri student achievement in Science.

Results presented below are based on Missouri student census data. The results presented here may differ slightly from the official state summary report of all student populations due to ongoing resolution of test materials and student information. The results in the tables in this chapter presented as evidence of reliability and validity of the scores from the Science MAP assessments and should not be used for state accountability purposes.

7.1. Cross-year, Cross-sectional Comparisons

It is often desirable to examine the scores of students across time. Because the test content remained the same for Science Grades 5 and 8 assessments between the 2008 and 2016 administrations, the 2016 test scores for Science are continued to be reported on the same scale as in the previous years. Table 7.1 shows state-level means for Grades 5 and 8 from 2008 through 2016. The mean scale score for Grade 5 decreased approximately 3 scale score points while the mean scale score for Grade 8 decreased less than 1 scale score point between the 2015 and 2016 test administrations.

Table 7.2 shows the percentage of students in each achievement level from 2008 through 2016 on the Science test. In Grade 5, the percentage of students at or above *Proficient* decreased by approximately 4% and in Grade 8, the percentage of students at or above *Proficient* decreased by less than 1% between the 2015 and 2016 test administrations. Because the same test forms were administered in 2015 and 2016 administration years, the performance of Grade 5 students on individual test items was examined in both administrations. It was observed that the 2015 Grade 5 cohort performed better on several constructed response items compared to the 2016 Grade 5 student population, which likely contributed to the overall lower performance on the 2016 Grade 5 Science assessment. The test items and scoring rubrics for these items did not change between the 2015 and 2016 test administrations.

It should be noted that invalidated students were assigned to the LOSS (Lowest Obtainable Scale Scores) and to the *Below Basic* achievement level. These students were excluded from the scale score summaries.

DRAFT

Table 7.1. Comparison of State-Level Means, 2008 through 2016: Science

Grade	Year	N	Mean SS	S.D. SS
5	2008	65,586	661.64	31.52
	2009	67,118	662.22	30.40
	2010	66,558	664.76	32.48
	2011	67,196	666.04	33.43
	2012	66,492	667.99	34.23
	2013	65,850	667.54	33.03
	2014	65,935	664.06	30.50
	2015	66,381	664.00	31.72
	2016	66,871	660.97	32.31
8	2008	67,209	694.36	30.67
	2009	66,702	695.65	30.94
	2010	66,101	698.28	31.07
	2011	65,828	700.05	30.98
	2012	66,724	700.18	31.92
	2013	66,418	699.92	31.71
	2014	66,912	701.94	29.53
	2015	66,430	698.19	31.67
	2016	65,730	697.50	32.64

DRAFT

Table 7.2. Comparison of Percentage of Students in each Achievement Level, Science 2008 through 2016 Census Data

Grade	Year	N	No Level	Below Basic	Basic	Proficient	Advanced	Prof & Adv
5	2008	65,734	0.2	11.2	44.0	29.6	14.9	44.5
	2009	67,307	0.3	10.6	44.1	30.3	14.8	45.1
	2010	66,730	0.3	10.4	40.5	29.6	19.3	48.9
	2011	67,461	0.4	10.0	39.1	29.5	21.0	50.5
	2012	66,675	0.3	9.8	38.5	27.2	24.3	51.4
	2013	65,980	0.2	9.6	39.0	28.1	23.1	51.3
	2014	66,153	0.4	9.0	43.3	31.5	15.9	47.3
	2015	66,411	0.0	10.6	42.4	28.8	18.1	47.0
	2016	66,914	0.1	11.8	45.2	28.0	14.9	42.9
8	2008	67,574	0.5	19.3	37.0	36.7	6.5	43.2
	2009	67,077	0.6	18.2	36.5	37.2	7.6	44.8
	2010	66,463	0.5	16.4	35.1	38.4	9.6	48.0
	2011	66,205	0.6	15.7	33.7	38.6	11.4	50.0
	2012	67,037	0.5	16.1	33.8	37.0	12.6	49.6
	2013	66,710	0.4	15.7	33.8	38.4	11.6	50.0
	2014	67,168	0.4	12.8	35.0	40.5	11.4	51.9
	2015	66,524	0.1	16.7	34.3	39.1	9.8	48.9
	2016	65,825	0.1	16.9	34.9	38.5	9.5	48.0

DRAFT

CHAPTER 10: FAIRNESS

As noted in the *Standards* (AERA, APA, & NCME, 2014), there are varying definitions of fairness. In this chapter, we examine fairness as it relates to minimizing bias on a test. We then look at test performance among varying subgroups assessed by Science MAP. It should be noted that differences in test performance among subgroups do not mean that a test is unfair—it simply means that groups perform differently on the test. Even when a test is carefully and properly constructed, differences may exist among subgroups as a result of differences in curriculum or learning by students in the subgroup.

10.1. Evaluating Bias through Differential Item Functioning (DIF) Statistics

After administering the test, an empirical approach known as differential item functioning (DIF) was used to examine the items. 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. The DIF procedures used and the results of these analyses are detailed in this section. It should be noted, though, that all items included in the Spring 2016 MAP have been thoroughly reviewed for content and bias by Missouri educators and DRC content experts ensure that they do not tap knowledge or specific ability irrelevant to the construct the test intends to measure. Therefore 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). Items are not necessarily suppressed from operational scoring if they are flagged for DIF.

DRAFT

The position of DRC 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 such biases, DRC strives to minimize the role of extraneous elements, thereby increasing the number of students for whom the test is appropriate. As discussed above, 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 follows (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 kth 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:

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

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:

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

For selected-response items, the MH (χ^2_{MH}) 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^2_{MH}, 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:

DRAFT

- 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 is 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 Science MAP raw score. SMD is computed as follows (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 kth 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 $|ES|$ is between 0.17 and 0.25.
- Large DIF: If the MH statistic is significant ($p < .05$) and $|ES| \geq 0.25$.

A positive DIF value indicates that the item favors the focal group, while a negative value indicates that the item disadvantages the focal group. Table 10.1 shows the DIF results for the following subgroups:

- ***Gender***: Focal group is females; reference group is males.

- ***Race/Ethnicity***: Focal groups are students whose race/ethnicity is reported as Black, Hispanic, Asian/Pacific Islander, American Indian/Alaska Native, or Other; reference group is students whose race/ethnicity is reported as White.

- ***Accommodations/Universal Tools***: Focal group is students who had one or more testing accommodations or universal tools indicated by a teacher; 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 minimum case count for the focal group was set at 200 and the minimum case count for the reference group was set at 400. The DIF analyses are not performed for subgroups of fewer than 200 students. In these cases, the statistical procedures do not have sufficient power to detect differences should they exist.

Table 10.1 summarizes the number of DIF flags by grade and test form for each focal group that included at least 200 students for Science assessments. For example, consider Grade 5 Science, form CA2 (see Table 10.1). In this form, two items were flagged for DIF for the female subgroup and one of them exhibited moderate negative DIF while the other one showed moderate positive DIF. Two items were flagged for DIF for the Asian/Pacific Islander subgroup. One of them

DRAFT

exhibited moderate negative DIF and one exhibited large negative DIF. Two items were flagged for the Black subgroup of which one display moderate negative DIF and one exhibited moderate positive DIF. One items was flagged against Hispanic group showing moderate negative DIF. Lastly, two items were flagged for the accommodated subgroup: one exhibited moderate positive DIF and one exhibited large positive DIF.

The number of American Indian/Alaska Native students taking each of the Science test form was smaller than 200 and no DIF was performed for this group.

Again, any items included on the Science MAP (including those items flagged for DIF) have been thoroughly reviewed for content and bias by Missouri teachers, DESE staff, and DRC Content experts and deemed appropriate for inclusion in the MAP assessments.

10.2. Evaluating Bias through Impact Analysis

The impact of achievement testing on minority subgroups can be determined and reported in the form of average scores and also in terms of test score reliability. Tables 10.2 and 10.3 present the number of students, scale score means and standard deviations, and effect size (Cohen's *d*) for the various subgroups of interest. Tables 10.4 and 10.5 show test form reliability statistics (coefficient alpha) for the different subgroups.

DRAFT

10.2.1. Effect Size

One way to evaluate the magnitude of the differences is to calculate the effect size. Cohen's *d* was used to calculate the effect size. Cohen's *d* is given by the formula

$$d = \frac{\overline{x_a} - \overline{x_b}}{\sqrt{\frac{(n_a - 1)s_a^2 + (n_b - 1)s_b^2}{(n_a + n_b) - 2}}},$$

where $\overline{x_a}$ is the mean score of group A, $\overline{x_b}$ is the mean score of group B, s_a^2 is the variance of group A, s_b^2 is the variance of group B, n_a is the number of students in group A, and n_b is the number of students in group B.

Cohen's *d,* then, expresses the difference in group means in terms of the standard deviation. For example if *d*=.34 for two groups, then it may be interpreted that the mean difference between the two groups is .34 of the pooled standard deviation. Cohen (1988) offers guidelines for interpreting the meaning of the *d* statistic: *d* = .20 is a small effect size, *d* = .50 is a medium effect size, and *d* = .80 is a large effect size.

Using Cohen's (1988) guidelines, certain trends become apparent in Tables 10.2 and 10.3. There is a small difference between the mean Science test scores of Asian/ Pacific Islander students and White students in Grade 8 with Asian/ Pacific Islander students outperforming White students. There is a large difference between the mean Science test scores of Black students compared to White students in both grades, where Black students underperform White students. There is a small difference between mean Science test scores of Hispanic students compared to White

students in Grade 5 and a medium difference in Grade 8, where Hispanic students underperform White students. There is a small difference between the mean Science test scores of American Indian/Alaska Native students compared to White students in both grades, where American Indian/Alaska Native students underperform White students. In addition, there is a medium difference between the mean Science test scores of students not needing testing accommodations or universal tools and students using testing accommodation or universal tools in both grades, where students not using testing accommodations or universal tools outperform students using testing accommodations or universal tools.

10.2.2. Reliability

Tables 10.4 and 10.5 show the regular test form reliability coefficients and SEM by student race/ethnicity, gender, and whether or not students are using any testing accommodations or universal tools for Grade 5 and Grade 8, respectively. The reliability coefficients ranged from 0.82 to 0.93. This analysis shows that the test reliability is of acceptable magnitude for all of the subgroups.

DRAFT

Table 10.1: 2015 MAP DIF Statistics: Number of Flagged Items, Science

Science		Grade	5		8	
		Form	CA2	CA3	CA2	CA3
		# of Items	41	41	39	38
Group	**DIF Magnitude**	**DIF Direction**	**Number of Flagged Items**			
Female	Moderate	Negative	1	1	2	0
		Positive	1	0	2	1
	Large	Negative	0	0	0	0
		Positive	0	0	1	1
Asian/ Pacific Islander	Moderate	Negative	1	2	1	0
		Positive	0	0	1	1
	Large	Negative	1	0	0	0
		Positive	0	0	0	0
Black	Moderate	Negative	1	1	1	0
		Positive	1	2	0	0
	Large	Negative	0	2	0	1
		Positive	0	0	0	0
Hispanic	Moderate	Negative	1	2	0	0
		Positive	0	0	0	0
	Large	Negative	0	0	0	0
		Positive	0	0	0	0
Ethnicity: Other	Moderate	Negative	0	0	0	0
		Positive	0	0	0	0
	Large	Negative	0	0	0	0
		Positive	0	0	0	0
Accommo-dations/ Universal Tools	Moderate	Negative	0	0	0	0
		Positive	1	0	0	0
	Large	Negative	0	0	0	0
		Positive	1	1	0	0

DRAFT

Table 10.2: Impact Analysis, Grade 5 Science

Category	Group	N	Mean	Std. Dev.	Effect Size
Race/ Ethnicity	White (not Hispanic)	48203	666.17	29.40	
	Asian/Pacific	1505	670.98	36.08	-0.16
	Black (not Hispanic)	10556	639.42	34.87	0.88
	Hispanic	4146	652.35	31.49	0.47
	Am. Indian/Alaska	276	657.30	30.00	0.30
	Other	2151	661.51	30.57	0.16
Gender	Male	34130	660.86	33.44	
	Female	32707	661.17	30.95	-0.01
Accommodations / Universal Tools	No	45276	668.55	28.41	
	Yes	21561	645.18	34.03	0.77

Table 10.3: Impact Analysis, Grade 8 Science

Category	Group	N	Mean	Std. Dev.	Effect Size
Race/ Ethnicity	White (not Hispanic)	47906	703.31	28.95	
	Asian/Pacific	1492	709.14	36.93	-0.20
	Black (not Hispanic)	10473	673.05	36.56	0.99
	Hispanic	3752	688.28	32.65	0.51
	Am. Indian/Alaska	320	692.99	31.60	0.36
	Other	1717	697.50	30.37	0.20
Gender	Male	33698	697.24	34.30	
	Female	31962	697.90	31.06	-0.02
Accommodations / Universal Tools	No	50839	703.01	29.10	
	Yes	14821	678.86	37.40	0.77

DRAFT

Table 10.4: Grade 5 Science Reliability and SEM by Subgroup

Form	Category	Group	N Count	Cronbach's Alpha	SEM
CA2	Race/Ethnicity	White (not Hispanic)	30544	0.86	3.58
		Asian/Pacific Islander	953	0.90	3.54
		Black (not Hispanic)	7510	0.88	3.57
		Hispanic	2925	0.87	3.62
		Am. Indian/Alaska N.	178	0.85	3.66
		Other	1354	0.87	3.60
	Gender	Male	22556	0.89	3.58
		Female	20908	0.88	3.59
	Accommodations/ Universal Tools	No	22908	0.85	3.56
		Yes	20556	0.88	3.58
CA3	Race/Ethnicity	White (not Hispanic)	17656	0.82	3.46
		Asian/Pacific Islander	552	0.85	3.32
		Black (not Hispanic)	3046	0.88	3.67
		Hispanic	1221	0.84	3.58
		Am. Indian/Alaska N.	98	0.86	3.52
		Other	797	0.84	3.52
	Gender	Male	11570	0.86	3.49
		Female	11798	0.85	3.52
	Accommodations/ Universal Tools	No	22364	0.84	3.50
		Yes	1004	0.89	3.65

DRAFT

Table 10.5: Grade 8 Science Reliability and SEM by Subgroup

Form	Category	Group	N Count	Cronbach's Alpha	SEM
CA2	Race/Ethnicity	White (not Hispanic)	28524	0.90	3.79
		Asian/Pacific Islander	878	0.93	3.71
		Black (not Hispanic)	6701	0.90	3.63
		Hispanic	2365	0.90	3.77
		Am. Indian/Alaska N.	195	0.90	3.81
		Other	1009	0.90	3.81
	Gender	Male	20720	0.92	3.74
		Female	18952	0.90	3.80
	Accommodations/ Universal Tools	No	25696	0.89	3.79
		Yes	13978	0.91	3.70
CA3	Race/Ethnicity	White (not Hispanic)	19376	0.87	3.76
		Asian/Pacific Islander	614	0.89	3.61
		Black (not Hispanic)	3772	0.88	3.65
		Hispanic	1387	0.88	3.79
		Am. Indian/Alaska N.	125	0.87	3.76
		Other	708	0.88	3.77
	Gender	Male	12978	0.89	3.73
		Female	13006	0.88	3.78
	Accommodations/ Universal Tools	No	25140	0.88	3.77
		Yes	842	0.89	3.58

DRAFT

References

Altman, D.G. (1991). *Practical statistics for medical research*. London: Chapman and Hall.

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

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

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.

Landis, J.R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159–174.

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.

DRAFT