Judging the Content and Statistical Equivalence of MCAS Operational and Linking Items[1,2]

Nina Deng, Tia Sukin, and Ronald K. Hambleton

University of Massachusetts Amherst

May 25, 2009

[1] Center for Educational Assessment MCAS Validity Report No. 20. (CEA-709). Amherst, MA: University of Massachusetts, Center for Educational Assessment.

[2] This research was carried out for the Massachusetts Department of Elementary and Secondary Education (MDESE) under a contract between the Center for Educational Assessment at the University of Massachusetts Amherst (Ronald Hambleton, PI) and Measured Progress (MP), Dover, NH.

Background

Educational test practitioners realize the need for ensuring a quality process for equating two or more forms of a test. Typically, when referring to criterion-referenced state-achievement exams such as the MCAS tests, examinees who take a particular test form are considered to be a part of a naturally occurring group. This means that the examinees are not randomly selected from some specified population, but instead, groups of examinees are formed based on the requirements of testing. For example, all 10^{th} graders in Massachusetts must take the mathematics test of the Massachusetts Comprehensive Assessment System (MCAS). Therefore, it is unlikely that examinees taking the test from one year to the next are equivalent. Especially with the desire to meet Annual Yearly Progress (AYP) goals, it is likely that the group of examinees taking later forms of the test have higher mean proficiency due to curricular and instructional modifications intended to better prepare students in all areas of interest. Non-equivalent groups from year to year make equating test forms challenging. Additionally, the test forms are rarely strictly parallel resulting in one form being more difficult than another. Differences between examinee groups and test forms must be taken into account when equating test forms.

The most common test equating design used to account and adjust for these differences is the Non-Equivalent Anchor Test (NEAT) design. This well-known design consists of using a set of items that appear on each of the tests to be equated and are often called anchor or common items (though not in Massachusetts). In the context of classical test theory (CTT), the anchor items are used to adjust for proficiency differences in the naturally occurring groups of examinees (e.g., Angoff, 1968; Angoff, 1971; Gulliksen, 1950; Holland & Dorans, 2006; Kolen & Brennan, 2004; Petersen, Kolen, & Hoover, 1989). However, with the MCAS, because of the Item Response Theory (IRT) focus, we will report on the use of anchor items to place item parameter estimates and ultimately proficiency scores from year to year onto the same scale (e.g., Hambleton, Swaminathan, & Rogers, 1991; Holland & Dorans, 2006; Kolen & Brennan, 2004; Lord, 1980). The content and statistical characteristics of the anchor items can have an impact on the quality of equating, and that will be the focus of this study.

Purpose of the Study

Until the 2008 MCAS technical report, the third author had never seen a study reported in any state that had looked at the content match between the linking items and the operational test items. We discussed the situation with a colleague, Professor Michael Kolen, an expert on equating (see Kolen & Brennan, 2006), and he indicated too that he had not seen reports from any states on the topic of this study. Content match between the anchor items and the operational items seems especially important, because it is in the linking process that growth from one year to the next is determined, and if content match (i.e., content representativeness) is not present, it would be possible to obtain a biased estimate of growth) and undermine the equating process. Of the two factors, content match is more important because non-match can lead to bias in estimates of growth. Statistical non-match can be handled fairly easily in the equating process.

The purpose of this study, therefore, was to investigate both the extent of the content match and the statistical match between the linking items and the operational MCAS test items. The study was carried out for all MCAS tests administered in 2008 that used fixed common item parameter estimation (grades 3, 4, 5, 6, 7, and 8 ELA and Mathematics, and grades 5 and 8 science and technology/engineering (STE). These are the tests where the match is of special interest. In the pre-equating design used at the Grade 10 level and with the high school science tests, the issue does not arise.

In the next section, we will provide a brief review of some of the relevant literature on linking items. What follows is a description of the sources of data we used to complete the study, followed by the results themselves. In a last section we have offered some conclusions and a suggestion for follow-up research.

Literature Review

Three important features of an anchor test are length, content, and statistical properties. Each of these features will be described in detail, including why it is important, relevant research regarding the feature, how it is to be evaluated, and possible negative consequences for straying from the commonly suggested guidelines. Table 1 on pages 12 and 13 helps to summarize the current recommendations in the field.

It is well known that the length of a single test is highly correlated with its reliability. Longer tests are often more reliable than shorter tests that measure the same construct (Angoff, 1968). Therefore, it would also be reasonable to assume that the length of an anchor test should be such that its reliability is of a respectable level. Table 1 helps summarize the recommendations found in the research literature for how long anchor tests should be in comparison to the operational test or how many items are required for placing item parameters on the same common metric. It is interesting to note that most the research suggests that the anchor test should represent at least 20% of the test or for IRT equating methods, at least 15 items should be used. After conducting empirical studies using IRT equating, Fitzpatrick (2008) found that shorter anchor test lengths seriously compromised the integrity of the equating results. She suggests this is due to the lack of parameter invariance associated with the close tie of tests to instruction resulting in shifts in parameters as instructional emphasis changes as noted by Cook, Eignor, and Taft (1988). She also notes that there seems to be lack of precision in anchor tests currently in use and proposes that instead of lengthening the anchor tests, we should consider investigating optimal allocation procedures familiar to the field of survey sampling (Sudman, 1976). This allocation method is based on sampling more elements from strata with more sampling variability. In the context of sampling items for inclusion in the anchor test, items from subsets known to have more variability based on content or statistical properties would be chosen in larger proportions than subsets with less variability given these features.

All of the relevant and available documentation speaks to the importance of creating an anchor test that proportionally follows the content specifications of the operational test when using the NEAT design (e.g., Cook, & Eignor, 1991; Cook, & Petersen, 1987; Dorans, Kubiak, & Melican, 1998; Hambleton, Swaminathan, & Rogers, 1991; Klein, & Jarjoura, 1985; Kolen, 1988; Kolen, & Brennan, 2004; Petersen, Kolen, & Hoover, 1989; Petersen, Marco, & Stewart,

1982; Sinharay, & Holland, 2006, 2007, 2008). If a content area is omitted, over-represented, or under-represented and growth occurs in this area, the amount of overall growth for the construct being measured may be incorrectly estimated. Therefore, the linking of tests will be incorrect as any change that takes place over time is reflected only in the anchor items. Klein and Jarjoura (1985) compared a content representative anchor to a long anchor without content representation and found that the shorter anchor with content representation proved to perform best when using two classical test theory methods for equating: Tucker linear equating and Levine equating. This is the study often cited supporting the recommendation that anchor and operational tests contain equivalent proportions of items representing differing content areas. Additionally, Yang (2000) studied four anchor item sampling designs and four equating methods, two of which utilized IRT designs. Under all methods of equating, Yang (2000) found that equating accuracy was greatest when using the item-sampling scheme that selected items for inclusion in the anchor test in such a way that the anchor items proportionally matched the content specifications for the entire test.

Similar to the research related to content matching, recommendations from the literature supports that the anchor test be composed of items that mimic the statistical properties of the operational test (e.g., Angoff, 1968; Cook, & Eignor, 1991; Dorans, Kubiak, & Melican, 1998; Kolen, 1988; Kolen, & Brennan, 2004; Petersen, Marco, & Stewart, 1982; Petersen, Kolen, & Hoover, 1989; Petersen, Marco, & Stewart, 1982). Often, this is called a "mini-test" and consists of items with similar mean difficulty and similar range of difficulty. As an example, Petersen, Marco, and Stewart (1982) studied several equating methods (external vs. internal, content similarity, mean difficulty similarity) with a total of 11 conditions using the SAT (Scholastic Aptitude Test) and TSWE (Test of Standard Written English) and found that matching the mean difficulty of test and anchor test items was a more important factor in ensuring a reliable anchor test for equating test forms when compared to the scenario where moderate differences in the content of anchor and operational tests existed. This conclusion was based on equating a test to itself via equipercentile methods. Additionally, they found that the mini-test performs best as an anchor overall and that equipercentile equating performs better than linear equating when differences in difficulty do exist between the anchor and operational test forms.

While Sinharay and Holland (2006, 2007, 2008) support the "mini-test" configuration when using an internal anchor test[3] design, they do not fully support these ideas when considering the external anchor test[4]. These authors present evidence supporting the semi-midi and midi-test forms as anchors instead of the mini-test form where the spread of the item difficulties are more constrained in order to preserve items that are either very easy or very hard in terms of the item's estimated difficulty parameter. These authors found that such anchor tests perform equally well and sometimes better than the mini-test when using post-stratification and equipercentile equating methods. Additionally, the midi- and semi-midi anchor tests were found to have higher anchor-test-to-total-test correlations than the minitests. Further work must be

[3] Internal anchor test items consist of items that are included in operational forms being equated and are used in the calculation of proficiency scores.

[4]External anchor test items consist of items included in operational forms being equated but they are not used in the calculation of proficiency scores. External anchor test items are used with the MCAS tests.

conducted using IRT equating methods to generalize these results. The suspicion is that the more multidimensional a test is, the less likely the Sinharay and Holland findings might be.

Test practitioners are encouraged to look at the differences in content and statistical properties between their operational and anchor tests. Doing so, requires constructing tables that explicitly show the percentage of points allocated to each content area for the operational and anchor test. For example, the grade 7 English Language Arts subtest of the MCAS is made up of three strands and four item types. Each of these strands and item types should be tabled to determine the appropriateness of the match between the operational and anchor test. If there is a noticeable difference, the anchor test should be reconstructed or consequences of these differences should be systematically investigated. Such an investigation would involve testing for the error associated with item sampling and constructing confidence intervals around the ability score estimates for each content area obtained during the linking process. Practical consequences due to these differences could be determined by investigating whether examinees are misclassified into proficiency levels due to the discrepancies between the operational and anchor tests. Item parameter estimates between the operational and anchor test can be evaluated in a similar manner. Here, the mean and standard deviations of the item parameter estimates would be tabulated by item type for each the operational and anchor test. Standard errors for equating (SEE) should be reported as described by Holland and Dorans (2006) in conjunction with practical measures of the difference that matters (DTM) described by Dorans and Feigenbaum (1994).

Methodology

The study was straightforward to carry out, once we found the data we needed. The tables of content distribution for the operational test and the anchor test provide the score points of the items falling in each content area and in each item format type, separately for the operational items and the anchor items. The numbers were derived from the item list files provided by Measured Progress, which indicate the item format and whether the item is an operational item or a linking item, and also from the content files which indicate which content area each item belongs to. The score is 1 point for each multiple-choice question (MCQ) or short- answer item, 4 for each constructed response (CR) item (except for math grade 3 in which the maximum is 2 points for each CR item), and 20 for each writing item. The percentage of score points in each content strand were calculated and compared with the goals specified in the test blueprint (to address content validity) and more importantly for our purposes the percentage of score points in each content strand were compared between the operational test and the set of linking items.

The tables of item statistics for the operational test and the set of anchor test items provide the mean and standard deviations of the a-, b-, and c-parameter estimates for the MCQ items, and a- and b-parameters for the short answer and polytomous items, separately for the operational items and the anchor items. The item parameter estimates for the anchor items were provided in folder “PAR FROM LAST YEAR” by Measured Progress. The item parameter estimates for operational items were calibrated using the Fixed Common Item Parameter equating method using PARSCALE v4.1.

Results

For each 2008 MCAS test, we have provided two tables. The first table in the pair, provides the matching results regarding the content. In ELA, there are only two strands and these are crossed with three possible item formats (multiple-choice, short-answer, and constructed-response). In grades 4 and 7, an additional item format is used: the writing prompt. In mathematics, the items are classified into five strands, and in STE, four strands. The interesting comparisons are between the percent of operational and the anchor test score points in each strand.

Here is a summary of the main findings regarding the content match:

Table 2, Grade 3, ELA: A shift of four MCQs in the anchor from reading and literature (LT) to language (LA) would be needed to balance the content.

Table 4, Grade 4, ELA: Excluding the writing prompt from the comparison, the match is perfect.

Table 6, Grade 5, ELA: A shift of four MCQs in the anchor from LT to LA would be needed to balance the content.

Table 8, Grade 6, ELA: A shift of two MCQs in the anchor from LT to LA would be needed to balance the content.

Table 10, Grade 7, ELA: A shift of four MCQs in the anchor from LT to LA would be needed to balance the content.

Table 12, Grade 8, ELA: A shift of one MCQ in the anchor from LT to LA would be needed to balance the content.

Table 14, Grade 3, Math: A shift of about two MCQs in the anchor from measurement (ME) to number sense and operations (NS) would be needed to balance the content.

Table 16, Grade 4, Math: A shift of one or two MCQs in the anchor would be needed to balance the content.

Table 18, Grade 5, Math: The content match is perfect.

Table 20, Grade 6, Math: The content match is perfect.

Table 22, Grade 7, Math: A shift of probably one MCQ in the anchor would be needed to balance the content.

Table 24, Grade 8, Math: A shift of several points, perhaps three or four might be needed to balance the content.

Table 26, Grade 5, Science: Probably the movement of a single MCQ item from life sciences (LS) to earth sciences (ES) in the anchor would be sufficient to balance the content.

Table 28, Grade 8 Science: The content balance is perfect.

It is also interesting to compare the desired distribution of content across the content strands as laid out in the content specifications and the actual distribution in the operational test. This comparison addresses the content validity of each test. Here, the comparisons show almost a perfect match, suggesting that content validity of the current tests (at the strand level) remains high. This finding was also confirmed in studies on the MCAS data carried out by Hambleton and Zhao a few years ago.

Tables 3, 5, 7, through Table 29 provide a statistical comparison of item statistics in the operational and linking sets. The statistics in the tables are reported separately for MCQ and polytomously scored items. In interpreting the differences, effect sizes were approximated and interpreted as follows: Small, less than about .10; Modest, around .25; Moderate, around .50; Substantial, around 1.00. Here is a summary:

Table 3, Grade 3, ELA: MCQ differences seem to be modest (relative to the variability in the b value statistics). Polytomous item differences are substantial.

Table 5, Grade 4, ELA: MCQ differences are substantial; poly item differences are more moderate.

Table 7, Grade 5, ELA: Both MCQ and poly item differences are modest.

Table 9, Grade 6, ELA: MCQ differences are moderate; poly item differences are small.

Table 11, Grade 7, ELA: Both MCQ and poly item differences are moderate.

Table 13, Grade 8, ELA: MCQ differences are modest; poly item differences are small.

Table 15, Grade 3, Math: MCQ and SA differences are moderate; poly item differences are substantial.

Table 17, Grade 4, Math: Both MCQ and poly item differences are substantial; SA differences are modest.

Table 19, Grade 5, Math: Both MCQ and poly item differences are modest; SA differences are moderate.

Table 21, Grade 6, Math: MCQ differences are modest; SA differences are moderate; poly item differences are small.

Table 23, Grade 7, Math: MCQ and SA item differences are small; poly item differences are modest.

Table 25, Grade 8, Math: MCQ differences are modest; poly and SA item differences are moderate.

Table 27, Grade 5, Science: Both MCQ and poly item differences are small.

Table 29, Grade 8, Science: MCQ differences are small; poly item differences are moderate.

Conclusions

The more important of the two matching variables is content, and here, only minor modifications would seem to be in order. With typically about 50 scoring points/test in the anchor, rarely do more than a couple of points (or an average of about two points per test or 4%) need to be adjusted. It is highly unlikely that adjustments as small as these can have much consequence on the equating results. Also, it must be noted that not all changes, however desirable, can be made, since there is not an unlimited supply of linking items assessing the content that is needed to match up the content on the linking and operational tests. And, any improvement to match up the content could have a negative impact on the match of item statistics.

As for the matching of item statistics in the operational tests and linking item sets, there seem to be more differences than we observed with the content match, but the differences are not huge and can easily be corrected for in statistical work associated with the equating. (Unfortunately no such match is possible with the content.) We note too that Sindaray and Holland even questioned the importance of matching the statistics. Still, to the extent possible, and until Sandaray and Holland's work can be validated, we would certainly recommend that some consideration in the selection of linking items be given to trying to match up the item statistics more carefully. But, we recognize that there are not an unlimited number of items than can be used. Given that a choice must be made, content would be the more consequential matching criterion.

While some work has been done to determine the number of items appropriate for calibrating item parameters within the context of a NEAT design and within an IRT framework, studies are lacking in determining how robust anchor sets are to discrepancies in content and statistical properties in comparison to their operational tests. Most of the research has focused on using CTT methods for equating. To date, the consensus seems to be that at least 15 items should be used to place item parameters on the same scale and the content of the anchor test should match that of the operational test. There is some question as to whether the anchor test must be a mini-version of the operational test with regard to statistical properties, such as the range of difficulty parameters. Some simulation research, or post-hoc analyses, along these general lines could be revealing.

References

Angoff, W. H. (1968). How we calibrate College Board scores. *College Board Review*, *68*, 11-14.

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement, 4th* ed. (pp. 508-600). Washington, DC: American Council on Education.

Budescu, D. (1985). Efficiency of linear equating as a function of the length of the anchor test. *Journal of Educational Measurement, 22*(1), 13-20.

Cook, L. L., & Eignor, D. R. (1989). Using item response theory in test score equating. *International Journal of Educational Research, 13*(2), 161-173.

Cook, L. L., & Eignor, D. R. (1991). IRT equating methods. *Educational Measurement: Issues and Practice, 10*, 37-45.

Cook, L. L., Eignor, D. R., & Taft, H. L. (1988). A comparative study of the effects of recency of instruction on the stability of IR and conventional parameter estimates. *Journal of Educational Measurement, 25*(1), 31-45.

Cook, L. L., & Petersen, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied Psychological Measurement, 11*(3), 225–244.

Dorans, N. J., & Feigenbaum, M. D. (1994). Equating issues engendered by changes to the SAT and PSAT/NMSQT. In I. M. Lawrence, N. J. Dorans, M. D. Feigenbaum, N. J. Feryok, A. P. Schmitt, & N. K. Wright (Eds.), *Technical issues related to the introduction of the new SAT and PSAT/NMSQT* (RM-94-10). Princeton, NJ: Educational Testing Service.

Dorans, N. J., Kubiak, A., & Melican, G. J. (1998). *Guidelines for selection of embedded common items for score equating* (ETS SR-98-02). Princeton, NJ: Educational Testing Service.

Fitzpatrick, A. R. (2008). NCME 2008 presidential address: The impact of anchor test configuration on student proficiency rates. *Educational Measurement: Issues and Practice, 27*(4), 34-40.

Gulliksen, H. (1950). *Theory of mental tests*. New York: John Wiley.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. Brennan (Ed.), *Educational measurement*, 4th ed. (pp. 187-220). Westport, CT: Praeger Publishers.

Klein, L. W., & Jarjoura, D. (1985). The importance of content representation for common-item equating with non-random groups. *Journal of Educational Measurement, 22*(3), 197-206.

Kolen, M. J. (1988). Traditional equating methodology. *Educational Measurement: Issues and Practice, 7*(4), 29-37.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling and linking: Methods and practices* (2nd ed.). New York: Springer.

Lord, F. M. (1980). *Practical applications item response theory*. Hillsdale, NJ: Lawrence Erlbaum.

McKinley, R. L., & Reckase, M. D. (1981). *A comparison of procedures for constructing large item pools* (Research Report 81-3). Columbia MO: University of Missouri, Department of Educational Psychology.

Michaelides, M. P., & Haertel, E. H. (2004). *Sampling of common items: An unrecognized source of error in test equating* (CSE Report 636). Los Angeles: Center for the Study of Evaluation (CSE), National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Graduate School of Education and Information Services, University of California at Los Angeles.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.; pp. 221-262). Washington, DC: American Council on Education.

Petersen, N. S., Marco, G. L., & Stewart, E. E. (1982). A test of the adequacy of linear score equating method. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 71-135). New York: Academic Press.

Sinharay, S., & Holland, P. W. (2006). *The correlation between the scores of a test and an anchor test* (ETS RR-06-04). Princeton, NJ: Educational Testing Service.

Sinharay, S., & Holland, P. W. (2007). Is it necessary to make anchor tests mini-versions of the tests being equated or can some restrictions be relaxed? *Journal of Educational Measurement, 44*(3), 249-275.

Sinharay, S., & Holland, P. W. (2008). Choice of anchor test in equating. *ETS Research Spotlight, 1*, 3-6.

Sudman, S. (1976). *Applied sampling*. New York: Academic Press.

Vale, C. D., Maurelli, V. A., Gialluca, K. A., Weiss, D. J., & Ree, M. J. (1981). Methods for linking item parameters (AFHRL-TR-81-10). Brooks Air Force Base TX: U.S. Air Force Human Resources Laboratory.

Wingersky, M. S., Cook, L. L., & Eignor, D. R. (1987). *Specifying the characteristics of linking items used for item response theory calibration* (Research Report 87-24). Princeton, NJ: Educational Testing Service.

Wingersky, M. S., & Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain procedures. *Applied Psychological Measurement, 8*, 347-364.

Table 1. Recommendations regarding anchor test composition based of length, content, and statistical properties.

Feature & Recommendation	Equating Framework	Research Conditions	Reference(s)
Length			
20 items or 20% of OT, which ever is larger	CTT	Rule of thumb, not based on research	Angoff, 1968, 1971
No clear recommendations	CTT	Derived an analytic solution for determining anchor test length for linear equating designs dependent upon the reliability of the operational test and the correlation of the anchor test to the operational tests	Budescu, 1985
Use as many as possible	IRT	Based on research reviewed	Cook, & Eignor, 1989
20% of OT	CTT/IRT	Rule of thumb, not based on research	Cook, & Eignor, 1991
At least 15 items	IRT	Based on simulated (Vale, et al., 1981) and empirical (Fitzpatrick, 2008; Reckase, et al.) studies	Fitzpatrick, 2008; Reckase, & McKinley, 1981; Vale, Maurelli, Gialluca, Weiss, & Ree, 1981;
20-25% of OT	IRT	Rule of thumb, not based on research	Hambleton, Swaminathan, & Rogers, 1991
20% of OT > 40 items or 30 items for very long tests		Based on experience and research reviewed	Kolen, & Brennan, 2004
As few as 2-5	IRT	Only advisable under the special circumstance where the two common items have very low standard errors for their parameter estimates and concurrent calibration methods are used	Wingersky, & Lord, 1984
20-40	IRT	Based on a simulation study using a 3PLM with concurrent and characteristic curve item calibration methods	Wingersky, Cook, & Eignor, 1987

Table 1. Recommendations regarding anchor test composition based of length, content, and statistical properties (continued).

Feature & Recommendation	Equating Framework	Research Conditions	Reference(s)
Content			
Anchor test must proportionally match the operational test's content specifications	CTT	Based on Klein, & Jarjoura's 1985 study using Tucker linear and Levine equating methods Kolen (1988) mathematically demonstrated this point via a hypothetical example	Cook, & Eignor, 1991; Cook, & Petersen, 1987; Hambleton, Swaminathan, & Rogers, 1991; Holland, & Dorans, 2006; Klein, & Jarjoura, 1985; Kolen, 1988; Kolen, & Brennan, 2004; Petersen, Kolen, & Hoover, 1989; Petersen, Marco, & Stewart, 1982; Sinharay, & Holland, 2006, 2007, 2008
Statistical Properties			
The anchor test should be a miniature version of the operational test in terms of the means and standard deviations of the item parameters	CTT/IRT	The CTT equating literature support this claim based on theory and studies such as Petersen et al. (1982) The IRT equating literature supporting this recommendation seem to be based on the CTT research	Angoff, 1968; Cook, & Eignor, 1991; Kolen, 1988; Kolen, & Brennan, 2004; Petersen, Marco, & Stewart, 1982; Petersen, Kolen, & Hoover, 1989; Petersen, Marco, & Stewart, 1982
The anchor test can be a semi-midi or midi-version of the operational test in terms of the standard deviation of the difficulty parameter	CTT	Recommended for external anchor tests only Found this to be true when using post-stratification and equipercentile equating methods	Holland, & Dorans, 2006; Sinharay, & Holland, 2006, 2007, 2008

Table 2. 2008 Grade 3 ELA Content Distribution for the Operational Test and the Anchor Test

Strand	Goal	Operational Test (Points = 48)				Anchor Test (Points = 48)			
		MCQ	SA	CR	%	MCQ	SA	CR	%
LA	15%	9	0	0	18.75	5	0	0	10.42
LT	82%	31	0	8	81.25	27	0	16	89.58

Percents are based on score points not test items. Numbers in the table are score points. Typically, the CR items are scored from 0 to 4.

Table 2. 2008 Grade 3 ELA Item Statistics for the Operational Test and the Anchor Test

Scoring	IRT Item Statistic	Statistic	Number of Operational Items	Operational Test	Number of Anchor Items	Anchor Test	Difference
0-1 Multiple Choice	b	$\bar{b}$	40	-0.964	32	-1.142	0.178
		SD(b)		0.613		0.674	-0.061
	a	$\bar{a}$		1.014		0.870	0.144
		SD(a)		0.207		0.234	-0.027
	c	$\bar{c}$		0.276		0.201	0.075
		SD(c)		0.090		0.071	0.019
Polytomous	b	$\bar{b}$	2	0.192	4	-0.521	0.713
		SD(b)		0.507		0.695	-0.188
	a	$\bar{a}$		0.562		0.716	-0.154
		SD(a)		0.083		0.154	-0.071

Table 4. 2008 Grade 4 ELA Content Distribution for the Operational Test and the Anchor Test

Strand	Goal	Operational Test (Points = 72)						Anchor Test (Points = 60)			
		MCQ	SA	CR	P	%	%(WP)	MCQ	SA	CR	%
LA	8%	6	0	0	0	8.33	11.54	7	0	0	11.67
LT	64%	30	0	16	0	63.89	88.46	29	0	24	88.33
Comp	28%	0	0	0	20	27.78	0.00	0	0	0	0.00

Percents are based on score points not test items. Numbers in the table are score points. Typically, the CR items are scored from 0 to 4. The prompt is worth 20 points and is only included on the grades 4, 7, and 10 ELA tests.

Table 5. 2008 Grade 4 ELA Item Statistics for the Operational Test and the Anchor Test

Scoring	IRT Item Statistic	Statistic	Number of Operational Items	Operational Test	Number of Anchor Items	Anchor Test	Difference
0-1 Multiple Choice	b	$\bar{b}$	36	-1.337	36	-0.883	-0.454
		SD(b)		0.494		0.548	-0.054
	a	$\bar{a}$		0.749		0.885	-0.136
		SD(a)		0.189		0.242	-0.053
	c	$\bar{c}$		0.185		0.203	-0.018
		SD(c)		0.065		0.059	0.006
Polytomous	b	$\bar{b}$	6	-0.352	6	0.086	-0.438
		SD(b)		0.874		0.303	0.571
	a	$\bar{a}$		0.904		0.901	0.003
		SD(a)		0.069		0.181	-0.112

Table 6. 2008 Grade 5 ELA Content Distribution for the Operational Test and the Anchor Test

Strand	Goal	Operational Test (Points = 52)				Anchor Test (Points = 60)			
		MCQ	SA	CR	%	MCQ	SA	CR	%
LA	12%	7	0	0	13.46	5	0	0	8.33
LT	88%	29	0	16	86.54	31	0	24	91.67

Percents are based on score points not test items. Numbers in the table are score points. Typically, the CR items are scored from 0 to 4.

Table 7. 2008 Grade 5 ELA Item Statistics for the Operational Test and the Anchor Test

Scoring	IRT Item Statistic	Statistic	Number of Operational Items	Operational Test	Number of Anchor Items	Anchor Test	Difference
0-1 Multiple Choice	b	$\bar{b}$	36	-1.096	36	-0.822	-0.274
		SD(b)		0.632		0.486	0.146
	a	$\bar{a}$		0.791		0.816	-0.025
		SD(a)		0.191		0.189	0.002
	c	$\bar{c}$		0.219		0.209	0.010
		SD(c)		0.075		0.074	0.001
Polytomous	b	$\bar{b}$	4	-0.182	6	-0.266	0.084
		SD(b)		0.270		0.279	-0.009
	a	$\bar{a}$		0.724		0.739	-0.015
		SD(a)		0.191		0.069	0.122

Table 8. 2008 Grade 6 ELA Content Distribution for the Operational Test and the Anchor Test

Strand	Goal	Operational Test (Points = 52)				Anchor Test (Points = 60)			
		MCQ	SA	CR	%	MCQ	SA	CR	%
LA	12%	5	0	0	9.62	4	0	0	6.67
LT	88%	31	0	16	90.38	32	0	24	93.33

Percents are based on score points not test items. Numbers in the table are score points. Typically, the CR items are scored from 0 to 4.

Table 9. 2008 Grade 6 ELA Item Statistics for the Operational Test and the Anchor Test

Scoring	IRT Item Statistic	Statistic	Number of Operational Items	Operational Test	Number of Anchor Items	Anchor Test	Difference
0-1 Multiple Choice	b	$\bar{b}$	36	-0.949	36	-0.613	-0.336
		SD(b)		0.549		0.698	-0.149
	a	$\bar{a}$		0.771		0.903	-0.132
		SD(a)		0.244		0.195	0.049
	c	$\bar{c}$		0.225		0.212	0.013
		SD(c)		0.074		0.066	0.008
Polytomous	b	$\bar{b}$	4	-0.314	6	-0.222	-0.092
		SD(b)		0.565		0.247	0.318
	a	$\bar{a}$		0.710		1.010	-0.3
		SD(a)		0.058		0.176	-0.118

Table 10. 2008 Grade 7 ELA Content Distribution for the Operational Test and the Anchor Test

Strand	Goal	Operational Test (Points = 72)						Anchor Test (Points = 59)			
		MCQ	SA	CR	P	%	%(WP)	MCQ	SA	CR	%
LA	8%	7	0	0	0	9.72	13.46%	4	0	0	6.78
LT	64%	29	0	16	0	62.50	86.54%	31	0	24	93.22
Comp	28%	0	0	0	20	27.78	0.00%	0	0	0	0.00

Percents are based on score points not test items. Numbers in the table are score points. Typically, the CR items are scored from 0 to 4. The prompt is worth 20 points and is only included on the grades 4, 7, and 10 ELA tests.

Table 11. 2008 Grade 7 ELA Item Statistics for the Operational Test and the Anchor Test

Scoring	IRT Item Statistic	Statistic	Number of Operational Items	Operational Test	Number of Anchor Items	Anchor Test	Difference
0-1 Multiple Choice	b	$\bar{b}$	36	-1.258	36	-0.948	-0.310
		SD(b)		0.624		0.458	0.166
	a	$\bar{a}$		0.810		0.797	0.013
		SD(a)		0.231		0.211	0.020
	c	$\bar{c}$		0.186		0.191	-0.005
		SD(c)		0.067		0.068	-0.001
Polytomous	b	$\bar{b}$	6	-0.587	6	-0.156	-0.431
		SD(b)		0.773		0.352	0.421
	a	$\bar{a}$		0.997		1.079	-0.082
		SD(a)		0.087		0.081	0.006

Table 12. 2008 Grade 8 ELA Content Distribution for the Operational Test and the Anchor Test

Strand	Goal	Operational Test (Points = 52)				Anchor Test (Points = 60)			
		MCQ	SA	CR	%	MCQ	SA	CR	%
LA	12%	6	0	0	11.54	6	0	0	10.00
LT	88%	30	0	16	88.46	30	0	24	90.00

Percents are based on score points not test items. Numbers in the table are score points. Typically, the CR items are scored from 0 to 4.

Table 13. 2008 Grade 8 ELA Item Statistics for the Operational Test and the Anchor Test

Scoring	IRT Item Statistic	Statistic	Number of Operational Items	Operational Test	Number of Anchor Items	Anchor Test	Difference
0-1 Multiple Choice	b	$\bar{b}$	36	-1.109	36	-0.849	-0.260
		SD(b)		0.634		0.499	0.135
	a	$\bar{a}$		0.821		0.928	-0.107
		SD(a)		0.231		0.256	-0.025
	c	$\bar{c}$		0.214		0.191	0.023
		SD(c)		0.057		0.053	0.004
Poly	b	$\bar{b}$	4	-0.615	6	-0.537	-0.078
		SD(b)		0.327		0.245	0.082
	a	$\bar{a}$		0.964		1.010	-0.046
		SD(a)		0.111		0.090	0.021

Table 14. 2008 Grade 3 Math Content Distribution for the Operational Test and the Anchor Test
(Linking Items 203467 and 207671 are not found in the parameter file from last year, therefore were deleted from Anchor test)

Strand	Goal	Operational Test (Points = 50)				Anchor Test (Points = 48)			
		MCQ	SA	CR	%	MCQ	SA	CR	%
GE	13%	3	0	4	12.50	2	1	4	13.16
ME	13%	1	2	4	12.50	3	1	4	15.79
NS	34%	10	2	4	35.00	9	1	4	31.58
PR	20%	6	0	4	20.00	5	1	4	21.05
SP	20%	5	1	4	20.00	4	1	4	18.42

Percents are based on score points not test items. Numbers in the table are score points. **The CR items are scored from 0 to 2.**

Table 15. 2008 Grade 3 Math Item Statistics for the Operational Test and the Anchor Test

Scoring	IRT Item Statistic	Statistic	Number of Operational Items	Operational Test	Number of Anchor Items	Anchor Test	Difference
0-1 Multiple Choice	b	$\bar{b}$	25	-1.014	23	-0.673	-0.341
		SD(b)		0.609		0.492	0.117
	a	$\bar{a}$		0.877		0.973	-0.096
		SD(a)		0.155		0.221	-0.066
	c	$\bar{c}$		0.218		0.184	0.034
		SD(c)		0.139		0.079	0.06
SA	b	$\bar{b}$	5	-1.186	5	-0.882	-0.304
		SD(b)		0.899		0.339	0.560
	a	$\bar{a}$		0.703		0.701	0.002
		SD(a)		0.283		0.205	0.078
Poly	b	$\bar{b}$	5	-0.241	5	-0.926	0.685
		SD(b)		0.524		0.559	-0.035
	a	$\bar{a}$		0.719		0.822	-0.103
		SD(a)		0.103		0.119	-0.016

Table 16. 2008 Grade 4 Math Content Distribution for the Operational Test and the Anchor Test

Strand	Goal	Operational Test (Points = 54)				Anchor Test (Points = 54)			
		MCQ	SA	CR	%	MCQ	SA	CR	%
GE	13%	4	0	4	14.81	2	1	4	12.96
ME	13%	1	1	4	11.11	3	0	4	12.96
NS	34%	13	2	4	35.19	12	2	4	33.33
PR	20%	6	1	4	20.37	7	1	4	22.22
SP	20%	5	1	4	18.52	5	1	4	18.52

Percents are based on score points not test items. Numbers in the table are score points. Typically, the CR items are scored from 0 to -4

Table 17. 2008 Grade 4 Math Item Statistics for the Operational Test and the Anchor Test

Scoring	IRT Item Statistic	Statistic	Number of Operational Items	Operational Test	Number of Anchor Items	Anchor Test	Difference
0-1 Multiple Choice	b	$\bar{b}$	29	-0.909	29	-0.190	-0.719
		SD(b)		0.891		0.774	0.117
	a	$\bar{a}$		0.858		1.010	-0.152
		SD(a)		0.197		0.273	-0.076
	c	$\bar{c}$		0.188		0.206	-0.018
		SD(c)		0.092		0.075	0.017
SA	b	$\bar{b}$	5	-0.069	5	-0.413	0.344
		SD(b)		1.146		0.885	0.261
	a	$\bar{a}$		0.772		0.764	0.008
		SD(a)		0.119		0.245	-0.126
Poly	b	$\bar{b}$	5	0.122	5	-0.397	0.519
		SD(b)		0.397		0.503	-0.106
	a	$\bar{a}$		0.828		0.877	-0.049
		SD(a)		0.158		0.179	-0.021

Table 18. 2008 Grade 5 Math Content Distribution for the Operational Test and the Anchor Test

Strand	Goal	Operational Test (Points = 54)				Anchor Test (Points = 54)			
		MCQ	SA	CR	%	MCQ	SA	CR	%
GE	13%	3	0	4	12.96	6	1	0	12.96
ME	13%	2	1	4	12.96	2	1	4	12.96
NS	33%	12	2	4	33.33	9	1	8	33.33
PR	26%	9	1	4	25.93	9	1	4	25.93
SP	15%	3	1	4	14.81	3	1	4	14.81

Percents are based on score points not test items. Numbers in the table are score points. Typically, the CR items are scored from 0 to 4.

Table 19. 2008 Grade 5 Math Item Statistics for the Operational Test and the Anchor Test

Scoring	IRT Item Statistic	Statistic	Number of Operational Items	Operational Test	Number of Anchor Items	Anchor Test	Difference
0-1 Multiple Choice	b	$\bar{b}$	29	-0.502	29	-0.529	0.027
		SD(b)		0.743		0.600	0.143
	a	$\bar{a}$		0.968		0.930	0.038
		SD(a)		0.261		0.244	0.017
	c	$\bar{c}$		0.182		0.207	-0.025
		SD(c)		0.086		0.066	0.02
SA	b	$\bar{b}$	5	-0.901	5	-0.447	-0.454
		SD(b)		0.716		0.533	0.183
	a	$\bar{a}$		0.728		0.787	-0.059
		SD(a)		0.113		0.052	0.061
Poly	b	$\bar{b}$	5	-0.135	5	0.051	-0.186
		SD(b)		0.544		0.395	0.149
	a	$\bar{a}$		0.928		1.043	-0.115
		SD(a)		0.150		0.038	0.112

Table 20. 2008 Grade 6 Math Content Distribution for the Operational Test and the Anchor Test

Strand	Goal	Operational Test (Points = 54)				Anchor Test (Points = 54)			
		MCQ	SA	CR	%	MCQ	SA	CR	%
GE	13%	2	1	4	12.96	2	1	4	12.96
ME	13%	3	0	4	12.96	2	1	4	12.96
NS	33%	12	2	4	33.33	13	1	4	33.33
PR	26%	9	1	4	25.93	9	1	4	25.93
SP	15%	3	1	4	14.81	3	1	4	14.81

Percents are based on score points not test items. Numbers in the table are score points. Typically, the CR items are scored from 0 to 4.

Table 21. 2008 Grade 6 Math Item Statistics for the Operational Test and the Anchor Test

Scoring	IRT Item Statistic	Statistic	Number of Operational Items	Operational Test	Number of Anchor Items	Anchor Test	Difference
0-1 Multiple Choice	b	$\bar{b}$	29	-0.518	29	-0.333	-0.185
		SD(b)		0.593		0.471	0.122
	a	$\bar{a}$		1.065		1.051	0.014
		SD(a)		0.236		0.252	-0.016
	c	$\bar{c}$		0.209		0.209	0.000
		SD(c)		0.082		0.082	0.000
SA	b	$\bar{b}$	5	-0.603	5	-0.949	0.346
		SD(b)		0.685		0.611	0.074
	a	$\bar{a}$		0.952		0.725	0.227
		SD(a)		0.207		0.189	0.018
Poly	b	$\bar{b}$	5	-0.289	5	-0.329	0.040
		SD(b)		0.482		0.593	-0.111
	a	$\bar{a}$		1.122		1.026	0.096
		SD(a)		0.127		0.197	-0.070

Table 22. 2008 Grade 7 Math Content Distribution for the Operational Test and the Anchor Test

Strand	Goal	Operational Test (Points = 54)				Anchor Test (Points = 55)			
		MCQ	SA	CR	%	MCQ	SA	CR	%
GE	13%	1	2	4	12.96	2	1	4	12.73
ME	13%	3	0	4	12.96	2	1	4	12.73
NS	26%	8	1	4	24.07	9	1	4	25.45
PR	28%	10	1	4	27.78	11	1	4	29.09
SP	20%	7	1	4	22.22	5	2	4	20.00

Percents are based on score points not test items. Numbers in the table are score points. Typically, the CR items are scored from 0 to 4.

Table 23. 2008 Grade 7 Math Item Statistics for the Operational Test and the Anchor Test

Scoring	IRT Item Statistic	Statistic	Number of Operational Items	Operational Test	Number of Anchor Items	Anchor Test	Difference
0-1 Multiple Choice	b	$\bar{b}$	29	-0.298	29	-0.237	-0.061
		SD(b)		0.718		0.590	0.128
	a	$\bar{a}$		0.983		1.165	-0.182
		SD(a)		0.212		0.244	-0.032
	c	$\bar{c}$		0.210		0.197	0.013
		SD(c)		0.074		0.072	0.002
SA	b	$\bar{b}$	5	-0.722	5	-0.729	0.007
		SD(b)		0.457		0.465	-0.008
	a	$\bar{a}$		0.852		0.828	0.024
		SD(a)		0.149		0.148	0.001
Poly	b	$\bar{b}$	5	-0.300	5	-0.181	-0.119
		SD(b)		0.515		0.374	0.141
	a	$\bar{a}$		1.194		1.172	0.022
		SD(a)		0.105		0.205	-0.100

Table 24. 2008 Grade 8 Math Content Distribution for the Operational Test and the Anchor Test

Strand	Goal	Operational Test (Points = 54)				Anchor Test (Points = 54)			
		MCQ	SA	CR	%	MCQ	SA	CR	%
GE	13%	1	2	4	12.96	3	1	4	14.81
ME	13%	3	0	4	12.96	5	1	4	18.52
NS	26%	9	1	4	25.93	5	1	4	18.52
PR	28%	10	1	4	27.78	9	1	4	25.93
SP	20%	6	1	4	20.37	7	1	4	22.22

Percents are based on score points not test items. Numbers in the table are score points. Typically, the CR items are scored from 0 to 4.

Table 25. 2008 Grade 8 Math Item Statistics for the Operational Test and the Anchor Test

Scoring	IRT Item Statistic	Statistic	Number of Operational Items	Operational Test	Number of Anchor Items	Anchor Test	Difference
0-1 Multiple Choice	b	$\bar{b}$	29	-0.123	29	0.098	-0.221
		SD(b)		0.659		0.842	-0.183
	a	$\bar{a}$		1.027		1.196	-0.169
		SD(a)		0.256		0.391	-0.135
	c	$\bar{c}$		0.208		0.220	-0.012
		SD(c)		0.112		0.074	0.038
SA	b	$\bar{b}$	5	-0.359	5	0.153	-0.512
		SD(b)		0.962		0.889	0.073
	a	$\bar{a}$		0.867		0.878	-0.011
		SD(a)		0.239		0.125	0.114
Poly	b	$\bar{b}$	5	-0.08	5	0.078	-0.158
		SD(b)		0.285		0.750	-0.465
	a	$\bar{a}$		1.192		1.208	-0.016
		SD(a)		0.245		0.212	0.033

Table 26. 2008 Grade 5 STE Content Distribution for the Operational Test and the Anchor Test

Strand	Goal	Operational Test (Points = 54)				Anchor Test (Points = 54)			
		MCQ	SA	CR	%	MCQ	SA	CR	%
ES	30%	9	0	8	31.48	12	0	4	29.63
LS	30%	11	0	4	27.78	12	0	4	29.63
PS	25%	10	0	4	25.93	6	0	8	25.93
TE	15%	4	0	4	14.81	4	0	4	14.81

Percents are based on score points not test items. Numbers in the table are score points. Typically, the CR items are scored from 0 to 4.

Table 27. 2008 Grade 5 STE, Item Statistics for the Operational Test and the Anchor Test

Scoring	IRT Item Statistic	Statistic	Number of Operational Items	Operational Test	Number of Anchor Items	Anchor Test	Difference
0-1 Multiple Choice	b	$\bar{b}$	34	-0.745	34	-0.682	-0.063
		SD(b)		0.806		0.707	0.099
	a	$\bar{a}$		0.779		0.785	-0.006
		SD(a)		0.232		0.159	0.073
	c	$\bar{c}$		0.267		0.240	0.027
		SD(c)		0.102		0.084	0.018
Polytomous	b	$\bar{b}$	5	-0.324	5	-0.443	0.119
		SD(b)		0.812		0.797	0.015
	a	$\bar{a}$		0.782		0.571	0.211
		SD(a)		0.134		0.106	0.028

Table 28. 2008 Grade 8 STE Content Distribution for the Operational Test and the Anchor Test

Strand	Goal	Operational Test (Points = 54)				Anchor Test (Points = 54)			
		MCQ	SA	CR	%	MCQ	SA	CR	%
ES	25%	10	0	4	25.93	10	0	4	25.93
LS	25%	10	0	4	25.93	10	0	4	25.93
PS	25%	9	0	4	24.07	5	0	8	24.07
TE	25%	5	0	8	24.07	9	0	4	24.07

Percents are based on score points not test items. Numbers in the table are score points. Typically, the CR items are scored from 0 to 4.

Table 29. 2008 Grade 8 STE Item Statistics for the Operational Test and the Anchor Test

Scoring	IRT Item Statistic	Statistic	Number of Operational Items	Operational Test	Number of Anchor Items	Anchor Test	Difference
0-1 Multiple Choice	b	$\bar{b}$	34	-0.156	34	-0.178	0.022
		SD(b)		1.009		0.652	0.357
	a	$\bar{a}$		0.936		0.825	0.111
		SD(a)		0.290		0.222	0.068
	c	$\bar{c}$		0.252		0.198	0.054
		SD(c)		0.080		0.070	0.010
Polytomous	b	$\bar{b}$	5	0.061	5	-0.165	0.226
		SD(b)		0.326		0.446	-0.120
	a	$\bar{a}$		1.039		1.019	0.020
		SD(a)		0.110		0.212	-0.102