UNIVERSITY OF MINNESOTA

Twin Cities Campus

Quantitative Methods in Education
Department of Educational Psychology
College of Education and Human Development

170 Education Sciences
56 East River Road
Minneapolis, MN 55455

612-624-4324
mcrdz@umn.edu

Memo To: Minnesota Assessment Group August 23, 2016

From: Michael Rodriguez

Subject Understanding Scores from the MCAs

Questions about meaningful and appropriate interpretation and use of MCA scores have been posed and discussed by MAG members recently and likely since the start of MAG.

To contribute to that effort, a closer look at the state MCA technical manual is informative, but since the material there is so dense, technical, and exhaustive, it is not easy to read – nor is it as direct regarding the appropriateness of test score use as many of us would prefer.

To provide deeper understanding of MCA test score interpretation and use, this brief review addresses the following:

- the meaning of criterion-referenced testing
- the underlying process of scoring and scaling the MCAs
- comments on computer adaptive testing
- interpretation guidelines and cautions provided by the state technical manual
- a more detailed illustration of why strand scores are not useful

A few initial implications from the technical manual can be made:

1. The MCA-III assessment system is designed to respond to federal school accountability requirements. As such, it is designed primarily to provide school-level information.
2. Scores are most appropriately used at the school-level, providing useful information regarding the distribution of performance of all students. They can be useful in identifying groups of students needing more support or evaluating the effects of programmatic changes and initiatives over time.
3. Student scores have measurement error that is estimated based on the idea of sampling items from the state standards-based content domains. Standard errors of measurement estimate how much a student's score might change if a different set of items were administered.
4. The measurement error that is reported with scores *does not* account for error in test scores due to different test settings or conditions, or performance over time, which include many of the ways in which we would like to interpret scores: "this score indicates how much a student knows and can do in the content area, not fixed to specific conditions at a specific time."
5. Content-strand scores are not useful (especially at the individual level), since they are based on so few items and are completely redundant with each other and the total score. They provide no unique information above and beyond the total score.

A Criterion-Referenced Testing System

The MCA-III[1] (hereafter simply referred to as MCA) is designed to be a series of criterion-referenced tests (CRTs), addressing federal educational accountability requirements to measure student performance relative to academic content standards in reading, mathematics, and science. The criterion-referenced aspect of the tests should provide a framework to make inferences about what students know and can do relative to the academic content standards.

- The "criterion" in CRT is the content domain –the referencing system for the MCAs is the content domain (some of us would prefer to use the term domain-referenced tests).
- To make CRTs effective, the content on the test must be a representative sample of the content domain – so that inferences can be made from a test score to the content domain.
- When test content does not represent the intended content domain, CRT inferences are limited. Some alignment evidence is available in reports by HumRRO for MN tests.

This is in contrast to norm-referenced tests (NRTs) where the referencing system is the norm-based distribution of scores, typically reported as percentiles. Percentiles can be reported in a CRT system, but do not convey criterion information, only the normative information.

- In NRTs, a student's score is referenced to the norming distribution of scores.
- Percentiles report how well a student performed compared to other students.

To support CRT interpretation, performance levels are provided and cut-scores are defined associated with each performance level. These performance levels are also required by federal regulations; they support score interpretation relative to the content domain. The presence or absence of performance levels and cut-scores does not make the MCAs a CRT – the CRT aspect of the test is due to our ability to make inferences about what students know and can do relative to the content domains, rather than make inferences about how a student performs relative to peers.

In addition, an important message is that students scoring near the cut score may be placed above or below the cut due to measurement error. Across students, these kinds of errors cancel out – making group distributions in performance levels more stable and more meaningful.

Measurement error is an indicator of the successful representation of the content domain:

- To the extent that the sample of items is a high quality representative sample of the domain, measurement error is minimized and scores are more consistent estimates of the domain knowledge, skills, and abilities of students.
- When the sample of items is poorly designed or small, more sampling error is reflected in the standard error of measurement (SEM).
- In CRTs, the classical SEM is a statistical estimate of the error due to sampling items.
- When we have more sampling error (larger SEM), our inferences to the content domain are less precise and less consistent (less reliable): if a student took a different sample of items, scores would likely change – the extent to which they might change is reflected in the SEM.

[1] Pearson. (2015). *Technical Manual for Minnesota's Title I and Title III Assessments for the Academic Year 2014-2015*. Roseville, MN: Minnesota Department of Education. Retrieved from http://education.state.mn.us/MDE/SchSup/TestAdmin/MNTests/TechRep/

MCA Scaling – Prior to Computer Adaptive Tests

The MCAs are scaled using a 3-parameter logistic (3PL) IRT model for selected-response items and generalized partial-credit (GPC) IRT model for constructed-response items. The 3PL model accounts for three aspects (parameters) of items:

- ***item difficulty*** – ability needed to have more than a 50% chance to get the item right;
- ***item discrimination*** – essentially the correlation between the item score and the total test score, which indicates the item's ability to represent the full domain as sampled in the test;
- ***lower-asymptote*** – the probability that the lowest ability students correctly answer an item, which may be due to guessing or to a base knowledge/ability that even the lowest ability students have (often incorrectly called the guessing parameter).

Raw-scores converted to IRT-scores are used to scale science MCAs, based on number-correct scores. This is achieved by finding the IRT ability score associated with each number-correct score through the test characteristic function resulting from the IRT modeling.

- In science, there is a one-to-one correspondence between number correct and scaled score
- The same raw score from different items correct results in the same scaled score.

Pattern scoring is used to score mathematics and reading MCAs (called measurement model-based scoring in the tech manual).

- The score a student receives is determined by the pattern of correct and incorrect responses to the administered items – scores depend on which items a student answers correctly.
- Two students with the same number correct could achieve different scores depending on which items were answered correctly.
- Essentially, getting more difficult items correct leads to higher ability estimates and more discriminating items are weighted more in estimating abilities.

If things are working well, pattern scores tend to be slightly more precise.

MCA Test and Score Quality

A personal observation: The MCA tests and resulting scores are as good as they get. The quality of test design, administration, and scaling meets industry standards and closely follows the *Standards for Educational and Psychological Testing*. The measurement error resulting from the MCAs is no larger, and in many cases, smaller than measurement error resulting from some of the most sophisticated large-scale standardized tests, including NAEP, the ACT, SAT, GRE, and others.

MCA Scaled Scores

To support score interpretation, a common metric was identified across grades and subjects, so that scores range from 1 to 99, with grade as a prefix (e.g., 301 to 399 for 3rd grade). Performance levels were also fixed on this scale, so that for each grade and subject matter:

- Partially Meets Standards cut score is fixed at G40.
- Meets Standards cut score is fixed at G50.
- Exceeds Standards cut scores vary across grades, ranging from G60 to G74.
- These scaled cut scores do not change from year to year.
- Cut scores are actually defined based on the ability required to achieve each performance level – as estimated by the IRT scaling, which does not change over time.
- What does change is the number correct associated with each ability. The number correct associated with each IRT-score changes as a function of the different items administered over time. But the ability required to meet standards is fixed and does not change.
- Scores are then equated over time to be placed on the same scale so that scores are comparable over time, within a subject and grade.

MCA Scaling

The spread of the score scale is defined by the distance between the Partially Meets and Meets Standards cut scores. The 10 point difference between 40 and 50 is divided by the difference in the cut-score thetas associate with the abilities at each cut score. The theta values, θs, are the ability estimates from the IRT scaling.

$$Spread = \frac{50 - 40}{(\theta_{Meets} - \theta_{PartiallyMeets})} = \frac{10}{(\theta_{Meets} - \theta_{PartiallyMeets})}$$

This spread is used to compute the scale score for each ability (θ) from the pattern scoring theta ability estimates in reading and math or thetas converted from number-correct scores in science.

$$Scale\ Score = (\theta - \theta_{Meets}) \times Spread + 50 + (Grade)(100)$$

In this formula, θ is the ability estimate based on student test performance.

- Ability estimates are first equated to the 2011 scale for grades 3-8 mathematics, 2014 for grade 11 mathematics, 2012 for science, 2013 for reading.
- To complete the equating, linking items are used to place new items (and ability estimates) on the previous scale, where linking items have known parameters based on the previous scale metric.
- Science scores undergo additional transformation because of the number-correct scoring.

Minimum and maximum scores are fixed.

- The lowest observable scale score (LOSS) is set to G01 and the highest observable scale score (HOSS) is set to G99.
- For mathematics and reading, because of constraints in the IRT scaling, in some grades, the G01 or G99 scores may not be observed.
- For science, in all cases, a score of zero correct is given the score G01; a score of all items correct is given the score G99. Scores with at least one wrong response will always be given values less than the HOSS.

Strand scores are based on a transformation of IRT scores estimated for each strand.

- A theta (θ) IRT score is estimated from the small number of items in each strand.
- The θ is transformed to the strand score scale, which ranges from 1-9.
- The SEM for strand scores ranges from 1 to 2 points.

$$\text{Strand Score} = 5 + \text{Round}(2 \times \theta)$$

Mathematics *Scaled Scores & SEM @ Cut-Scores and Lowest/Highest Scaled Scores*

	LOSS		Partial		Meets		Exceeds		HOSS	
Grade	SS	SEM	SS	SEM	SS	SEM	SS	SEM	SS	SEM
3	315	8.0	340	3.3	350	3.0	366	3.0	399	9.9
4	409	9.9	440	4.0	450	3.5	466	3.0	499	10.0
5	515	8.3	540	3.0	550	2.1	563	2.7	586	7.2
6	611	10.2	640	3.0	650	2.7	662	2.2	688	7.3
7	718	9.9	740	3.0	750	2.0	760	2.0	782	5.4
8	813	11.4	840	3.1	850	3.0	861	2.0	888	5.7
11	1102	21.9	1140	4.0	1150	3.0	1164	4.0	1095	10.2

Reading *Scaled Scores & SEM @ Cut-Scores and Lowest/Highest Scaled Scores*

	LOSS		Partial		Meets		Exceeds		HOSS	
Grade	SS	SEM	SS	SEM	SS	SEM	SS	SEM	SS	SEM
3	301	21.3	340	5.0	350	5.0	374	6.0	399	13.8
4	411	14.6	440	4.0	450	4.0	466	5.0	490	11.7
5	517	12.9	540	4.0	550	3.6	567	4.1	591	10.5
6	606	14.7	640	4.0	650	4.0	667	5.0	699	13.2
7	703	14.5	740	4.0	750	4.0	767	5.0	798	13.5
8	802	14.4	840	5.0	850	4.1	867	5.0	898	12.8
10	1013	11.4	1040	4.0	1050	3.9	1064	4.0	1094	10.0

Science *Scaled Scores & SEM @ Cut-Scores and Lowest/Highest Scaled Scores*

	LOSS		Partial		Meets		Exceeds		HOSS	
Grade	SS	SEM	SS	SEM	SS	SEM	SS	SEM	SS	SEM
5	501	5.0	540	5.0	550	5.0	570	7.0	599	1.0
8	801	7.0	840	3.0	850	3.0	863	5.0	899	2.0
HS	1001	7.0	1040	4.0	1050	3.0	1063	3.0	1099	1.0

Note. Scale scores at each cut-score, LOSS, and HOSS are the same each year. The SEM varies slightly from 2014 to 2015; SEM varies as a function of student variability and test score reliability each year.

Computer Adaptive Testing

Technical documentation has not been publically released (as far as I can see), so this is largely based on preliminary information from public presentations and descriptions. At this point, reading and mathematics online tests are adaptive.

- The online reading test is partially adaptive at the passage level. A student's performance on the item set for a passage (or passages) determines the next passage and items to be administered;
- The online mathematics test is adaptive at the item level.
- The CAT system is essentially locating the ability of a student by administering items where the ability of the student matches the difficulty of the items (so that a student may be getting about 50% of those items correct).
- The current MCA test specifications determine the proportion of items in each strand to be administered in the CAT.
- Each student will take the same proportion of items in each strand for a given subject.
- Within a subject/grade test, student scores are on the same scale as prior tests, although they will take different items based on their estimated ability.
- All items in the CAT pool are scaled on the original MCA-III scale (based on new field-testing items and calibrating them with previous items), resulting in consistent scores for each student. These IRT scores are then converted to scaled scores, using the same transformation formula as above.
- If students respond in ways consistent with their ability (when paying attention, putting forth full effort, having the opportunity to learn the standards), test scores should be slightly more precise, particularly those scores further away from the mean score (compared to non-CAT scores).

Comparing MCA-II and MCA-III

Because content and performance standards changed between MCA-II and MCA-III test series, there is no way to compare (even relatively) the level of ability to achieve any of the performance levels (cut scores). These are not comparable.

Appropriate Test Score Interpretation & Use

This section is taken directly from the MCA Technical Manual, pages 70-71.

Appropriate Score Uses

The tests in the Minnesota Assessment System are designed primarily to determine school and district accountability related to the implementation of the Minnesota standards. They are summative measures of a student's performance in a subject at one point in time. They provide a snapshot of the student's overall achievement, not a detailed accounting of the student's understanding of specific content areas defined by the standards. Test scores from Minnesota assessments, when used appropriately, can provide a basis for making valid inferences about student performance. The following list outlines some of the ways the student scores can be used.

• *Reporting results to parents of individual students*

The information can help parents begin to understand their child's academic performance as related to the Minnesota standards.

• *Evaluating student scores for placement decisions*

The information can be used to suggest areas needing further evaluation of student performance. Results can also be used to focus resources and staff on a particular group of students who appear to be struggling with the Minnesota standards. Students may also exhibit strengths or deficits in strands or substrands measured on these tests. Because the strand and substrand scores are based on small numbers of items, the scores must be used in conjunction with other performance indicators to assist schools in making placement decisions, such as whether a student should take an improvement course or be placed in a gifted or talented program.

• *Evaluating programs, resources and staffing patterns*

Test scores can be a valuable tool for evaluating programs. For example, a school may use its scores to help evaluate the strengths and weaknesses of a particular academic program or curriculum in their school or district as it relates to the Minnesota standards.

[*Note.* Underlined emphases are mine.]

Interpreting Variability in Performance

Strive to explore and display variability of scores within and between schools. We learn a great deal about the challenges faced by teachers and schools by understanding the variability in performance of our students. We don't teach to or plan instruction to address the percent proficient or the average score. We teach and work with students that are more or less variable in achievement.

This is consistent as discussed with MAG in the past, and in the Guidance Document produced with the help of MAG available at http://www.edmeasurement.net/MAG.

Interpretation Cautions (from the Technical Manual, mostly from page 122)

- Scores on different subjects and different grades are not comparable.
- Score differences are not comparable across subjects or grades. A difference of 5 points on one subject in a grade will represent an ability difference that does not compare to a 5-point difference on other subjects or grades.
- Scores are not comparable across grades, nor are score differences.
- Scores over time (within the same content standards period) are comparable within subject and grade.
- Achievement levels can be compared more safely across subjects and grades – relative to the performance level descriptors for a given test/grade. Monitoring trends in percent in each performance level across subjects and grades can be important to schools monitoring performance over time.

Below, you will find tables of the lowest and highest possible scores in each subject and grade, and the cut scores for each performance level. With these scores, the standard errors of measurement (SEM) are reported. Notice that scores that are further from the middle of the score scale (G50) are less precise.

- Low scores and high scores should be interpreted with more caution.
- The lowest and highest scores are the least precise.

Monitoring Performance or Gaps over Time

Remember that all of our communities (within schools and within districts) are changing at different rates over time. We might want to make direct inferences of percent of students in performance levels within a grade over time or across grades over time. However, each year we have a different population of students within a grade – and as students move from one grade to the next, it is rarely true that the population of students does not change.

The ideal picture for monitoring achievement gaps over time is to use a Panel Design, identifying a fixed group of students that are in a school or district over time (where the membership of the panel does not change).

You will find more information about this in the Guidance Document: *Analyzing and Reporting Achievement Gaps Guidance for Minnesota Schools* (page 25).

Available at http://www.edmeasurement.net/MAG

Strand Scores

Based on current evidence (MCA technical reports and yearbooks), strand scores do not provide unique information above and beyond the total scores and are severely limited in their ability to be diagnostic or inform instruction.

When strand scores are not very highly correlated, it is generally acknowledged that they may provide unique information above and beyond the total score. We also know that a correlation can be *no larger* than the score reliabilities. When a correlations is as large as the score reliabilities, strand scores do not provide additional unique information beyond measurement error (reliabilities limit score correlations).For a deeper read on this topic, see NCME Instructional Module on Subscores (2011)[2]. Reliabilities of scores limit correlations, as shown by:

$$Correlation_{XY} \leq \sqrt{Reliability_X \cdot Reliability_Y}$$

Correlations can be corrected (disattenuated) for measurement error. To do so, we manipulate the above relation and divide the correlation by the corresponding reliabilities:

$$\frac{Correlation_{XY}}{\sqrt{Reliability_X \cdot Reliability_Y}} \leq 1.0$$

The following tables contain original and corrected strand correlations in Reading and Mathematics.

- Corrected correlations are all near 1.0 – indicating complete redundancy of information.
- Strands do not provide unique or diagnostic information above and beyond the total score.

Corrected Strand Correlations for Reading 2015.

Grade		Reliabilities	Uncorrected Correlation	**Corrected Correlation**
3	Literature	.80		
	Information	.82	.81	1.00
4	Literature	.80		
	Information	.81	.80	.99
5	Literature	.81		
	Information	.79	.80	1.00
6	Literature	.78		
	Information	.83	.81	1.00
7	Literature	.79		
	Information	.83	.81	1.00
8	Literature	.80		
	Information	.82	.80	.99
10	Literature	.77		
	Information	.85	.80	.99

Source: Yearbook Tables 2014-15.

[2] https://www.ncme.org/ Click on [Publication] then [Items] on the left column. The 2011 Instructional Module can be found in the list near the bottom (ordered chronologically).

Corrected Strand Correlations for Mathematics 2015

Grade 3

Uncorrected	N&O	Alg	G&M	DA
N&O	*.85*			
Alg	.74	*.67*		
G&M	.81	.70	*.78*	
DA	.72	.62	.69	*.66*

Corrected	N&O	Alg	G&M
Alg	.98		
G&M	.99	.97	
DA	.96	.93	.96

Grade 4

Uncorrected	N&O	Alg	G&M	DA
N&O	*.87*			
Alg	.77	*.70*		
G&M	.75	.67	*.76*	
DA	.73	.65	.64	*.63*

Corrected	N&O	Alg	G&M
Alg	.99		
G&M	.92	.92	
DA	.99	.98	.92

Grade 5

Uncorrected	N&O	Alg	G&M	DA
N&O	*.86*			
Alg	.80	*.76*		
G&M	.73	.70	*.69*	
DA	.67	.64	.59	*.64*

Corrected	N&O	Alg	G&M
Alg	.99		
G&M	.95	.97	
DA	.90	.92	.89

Grade 6

Uncorrected	N&O	Alg	G&M	DA&P
N&O	*.85*			
Alg	.79	*.78*		
G&M	.79	.75	*.77*	
DA&P	.77	.72	.71	*.72*

Corrected	N&O	Alg	G&M
Alg	.97		
G&M	.98	.97	
DA	.98	.96	.95

Grade 7

Uncorrected	N&O	Alg	G&M	DA&P
N&O	*.81*			
Alg	.82	*.83*		
G&M	.75	.76	*.72*	
DA&P	.77	.78	.71	*.74*

Corrected	N&O	Alg	G&M
Alg	1.00		
G&M	.98	.98	
DA&P	.99	1.00	.97

Grade 8

Uncorrected	N&O	Alg	G&M	DA&P
N&O	*.67*			
Alg	.74	*.88*		
G&M	.64	.73	*.65*	
DA&P	.57	.69	.56	*.55*

Corrected	N&O	Alg	G&M
Alg	.96		
G&M	.97	.97	
DA&P	.94	.99	.94

Grade 11

Uncorrected	Alg	G&M	DA&P
Alg	*.87*		
G&M	.81	*.78*	
DA&P	.84	.71	*.69*

Corrected	Alg	G&M
G&M	.98	
DA&P	1.00	.97

Note. Scores on diagonals are reliabilities from online versions.
Source: Yearbook Tables 2014-15.