**M**innesota **A**ssessment **G**roup

**Principles of Statistics & Measurement**

**to Promote Teaching & Learning**

**(to Meet the Demands of Accountability)**

Michael C. Rodriguez

Educational Psychology, University of Minnesota

<http://www.edmeasurement.net/MAG>

Competency Standards in Student Assessment for Educational Administrators

This is available at the MAG resource website.

Educational Assessment Knowledge and Skills for Teachers

Brookhart, S.M. (2011). Educational assessment knowledge and skills for teachers. *Educational Measurement: Issues and Practice, 30*(1), 3-12.

A Statistical Basis

Variability

* Variability is an indicator that describes the degree of differences among individuals in terms of the measures we have obtained from those individuals.
* Our primary interest is in these differences: variability indicates the extent to which individuals differ.

Deviations

 Individuals differ from the mean.

Standard Deviation

 The typical deviation (difference)

Measurement

1904, Thorndike published ***An Introduction to the Theory of Mental and Social Measurements***

* Measurements are usually based on limited samples of behavior.
* The measurement obtained is always subject to error.

Classical Test Theory

*X* = *T* + *E*

Observed scores are the sum of the true score (T) plus measurement error (E); So are their variances (variability).

Reliability



A ratio of true to total variance; the proportion of variation that is true (real).

Standard Error of Measurement



SE Measurement = 

The part of the typical deviation that is due to measurement error (not true).

Interpreting Individual Scores

We observe a score, and due to sampling of items/time/occasions/settings/etc, we know an individual’s score will vary on average by the *SEM*, simply as a function of measurement error.

Validity First

* Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests.
* Evidence of the validity of a given interpretation of test scores for a specified use is a necessary condition for the justifiable use of the test.
* *2014 Standards for Educational & Psychological Testing*

Validity Issues

* Intended claims and uses
* Test score use (decisions) must be accompanied by additional data and relevant evidence
* No decision should be made from a single test score
* Validity is a matter of degree – more high-stakes decisions require stronger validity evidence

Validation

The process of validation involves accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretation.

*2014 Standards for Educational & Psychological Testing*

Sources of Validity Evidence

* Content is relevant, given OTL
* Cognitive process
* Internal (factor) structure
* Relations to other conceptually related constructs
* Relations to criteria or future performance
* Consequences interfere with use

Validity Related Issues

* Reliability is validity evidence, regarding score precision and generalizability of score interpretation and use
* Subscale meaning and independence
* Measurement invariance across subgroups

Reliability

* Reliability refers to the consistency of scores across replications of a testing procedure
* It is estimated or reported in many ways:
	+ Standard errors
	+ Reliability coefficients
	+ Generalizability coefficients
	+ Classification consistency

Reliability Related Issues

* Reliability/precision is always important.
* The need for precision increases as the consequences of decisions and interpretations grow in importance.

Validity assumes Reliability

* Reliability is a characteristic of test scores; the degree of consistency in test-taker scores over replications of a measurement procedure.
* The challenge, then, is to specify what constitutes a replication of the measurement procedure.
* Question: What is the relevant notion of replication for your use?

Measures of Internal Consistency

* Interested in error due to content sampling from domain or flawed items; interested in generalizability over items, forms, or tasks
* Estimated with internal consistency index associated with measurement model (coefficient alpha, the poorest index of all)
* Not necessarily a good thing for predictive validity since internal consistency is gained through item homogeneity – which limits correlations with broader behavioral outcomes

Measures of Equivalence

* Interested in error due to domain sampling error from form to form; interested in generalizability over forms
* Estimated with a correlation between two randomly-parallel forms

Measures of Stability

* Interested in error due to trait instability; interested in generalizability over occasions
* Estimated by test-retest correlation on the same test
* Time interval and relevant intervening experiences must be specified

Significant Limitations

Estimates of internal consistency are very sensitive to construct-irrelevant factors:

* Test length
* Group variability
* And do not capture random variation due to
* Time of year
* Context of testing
* Personal dispositions that vary over time

Applying Fairness Principles

* Fairness is a fundamental validity issue and requires attention throughout all stages of test development and use.
* Potential threats to score interpretation and use exist for individuals that differ in English proficiency (native and non-native English speakers), specific disabilities, cultural experiences, prior experience in the system.

Fairness

* Fairness applies to
* Treatment during the testing process
* Lack of measurement bias (score level, prediction, precision)
* Accessibility – opportunity to demonstrate standing on the measure
* Opportunity to learn the relevant content (with appropriate supports)
* Access to high-quality instruction
* Validity of test score interpretations and uses

Score Interpretation

* We are more interested in the likely stable value – the stable part of the score – not the unstable observed value.

**Lessons Learned about Testing**

A Report of the

 *National Research Council*

This is available at the MAG resource website.

*Re: Uses*

* In many situations, standardized tests provide the most objective way to compare the performance of a large group of examinees across places and times.
* A test score is an estimate rather than an exact measure of what a person knows and can do.
* High-stakes decisions about individuals should not be made on the basis of a single test score.
* Tests should not be used for high-stakes decisions if test takers have not had an opportunity to learn the material on which they will be tested.
* States, districts, and schools should aim to maximize the participation of English-language learners and students with disabilities in large-scale tests.
* Teachers need professional development that helps them better understand core principles of assessment and how to apply these to their regular instruction and testing.

*Re: Design*

* In the design of tests, form must follow function.
* The design process must ensure that test score interpretations are valid.
* The design process must ensure that the test results are reliable and fair.
* Testing professionals should consider the relationships among cognition, observation, and interpretation—the “assessment triangle”—when evaluating the soundness of current educational tests or designing new ones.
* Advances in cognitive sciences and measurement offer opportunities to develop educational assessments that better support learning.

*Re: Consequences*

* The people who design and mandate tests must be constantly vigilant about equity concerns, including opportunity to learn, cultural bias, or adverse impact.
* In the absence of effective services for low-performing students, better tests will not lead to better educational outcomes.
* Test results may be invalidated by teaching narrowly to a particular test.
* New testing programs should build in an evaluation component.

*Re: Public Understanding*

* Test developers and policy makers should clearly explain to the public the purpose for a test and the meaning of different levels of test performance.
* When test results are reported to students, teachers, and the public, the limitations of the test should be explained clearly to a lay audience.

Using Principles of Measurement *and Lessons Learned*

to Support Test Score Interpretation and Use in Minnesota

National Academies (2009)

A test score is an **estimate** rather than an exact measure of what a person knows and can do. The items on any test are a sample from some larger universe of knowledge and skills, and scores for individual students are affected by the particular questions included.

*Board of Testing and Assessment*

National Academies (2009)

A student may have done better or worse on a different sample of questions. In addition, guessing, motivation, momentary distractions, and other factors introduce uncertainty into individual scores.

Measurement Error ↔ Sampling

**2014-2015 Technical Manual for
Minnesota’s Title I and Title III Assessments**

***Understanding Measurement Error***

When interpreting test scores, it is important to remember that test scores contain some amount of measurement error. That is to say, test scores are not infallible measures of student characteristics… **measurement error must always be considered** when making score interpretations. (p. 73)

2014-2015 Technical Manual

***Using Objective/Strand-Level Information***

Strand or substrand level information can be useful as a preliminary survey to help identify skill areas in which further diagnosis is warranted. The standard error of measurement associated with these generally brief scales makes drawing inferences from them at the individual level very suspect; more confidence in inferences is gained when analyzing group averages. (p. 74)

2014-2015 Technical Manual

When considering data at the strand or substrand level, the error of measurement increases because the number of possible items is small. In order to provide comprehensive diagnostic data for each strand or substrand, the test would have to be prohibitively lengthened. (p. 74)

MCA for Individual Interpretation

2014-2015 Yearbook Tables for Minnesota’s Title I and Title III Assessments

Example: Grade 3 Reading

Score Distributions, p. 91



Seeking More Information

* Looking to Subscales for more info
* Knowing subscales are shorter – less precision
* Consider the role of measurement error in correlations
* Randomness doesn’t correlate with anything
* Measurement error (random noise) limits correlations

2015 MCA-III Summary Statistics
Grade 3 Reading (p. 134)



2015 MCA-III Subscale Correlations
Grade 3 Reading (p. 163)







3rd Grade Mathematics Subscale Corrected Correlations



6th Grade Mathematics Subscale Corrected Correlations



8th Grade Mathematics Subscale Corrected Correlations



2015 MCA Reading Subscale Corrected Correlations:

***Literature*** & ***Information***



National Academies 2009

The choice of appropriate assessments for use in instructional improvement systems is critical. Because of the extensive focus on large-scale, high-stakes, summative tests, policy makers and educators sometimes mistakenly believe that such tests are appropriate to use to provide rapid feedback to guide instruction. This is not the case.

National Academies 2009

Tests that mimic the structure of large-scale, high-stakes, summative tests, which lightly sample broad domains of content taught over an extended period of time, are unlikely to provide the kind of fine-grained, diagnostic information that teachers need to guide their day-to-day instructional decisions.

National Academies 2009

…BOTA urges the Department to clarify that assessments that simply reproduce the formats of large-scale, highstakes, summative tests are not sufficient for instructional improvement systems.

A Basis for Score Interpretation

*Criterion-Referenced Testing*

* The “criterion” in CRT is the content domain –the referencing system for the MCAs is the content domain.
* To make CRTs effective, the content on the test must be a representative sample of the domain – so 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.

In Contrast…

*Norm-Referenced Testing*

* In NRTs, a student’s score is referenced to the norming distribution of scores.

Performance Levels & Cutscores

To support CRT interpretation, performance levels are provided and cut-scores define each performance level. These performance levels are 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.

Supporting CRT Score Interpretation

* The test specifications (from the content standards framework) can be viewed as the sampling design, from which a sample of items is drawn to represent 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.

Supporting CRT Score Interpretation

* 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.

Supporting CRT Score Interpretation

* 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.

MCA Score Interpretation Cautions

* 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.

MCA Score Interpretation Cautions

* Achievement levels can more safely be compared across subjects and grades – rely on the performance level descriptors for a given test/grade.
* Scores over time (within the same content standards period) are comparable within subject and grade.

Appropriate Score Use

* 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.

Appropriate Score Use: Parents

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

Appropriate Score Use: 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.

Appropriate Score Use: Placement Decisions

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.

Appropriate Score Use: Program Evaluation

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.

A few *thoughts*

DDDM is a core element of P-12 school functioning.

“We live and breathe DDDM every day. It drives what we do and when we do it.”

From a study of Twin Cities educators and their thoughts on DDDM.

…*thoughts*

* More of the recent graduates arrive at school ready to engage in DDDM.
* The models vary widely:
	+ Data retreats
	+ Data wall
	+ Team data reviews
	+ Data exchanges

…*thoughts*

* The meaningfulness/usefulness of data and data use is often attributed to the level of training of school leaders – Principals matter.
* Where central school support for DDDM is less present and central leadership is not clear and consistent, DDDM was less present throughout the school – educators saw limitations in the decision-making activities.

…*thoughts*

* Accountability systems are designed to maximize effort, innovation, self-correction: optimal outcomes
* In many cases, they lead to maximizing self-justification, motivating options and decision paths that are easy to justify

…*thoughts*

* By simply increasing attention to our decision-making process, there is no guarantee that new ways of solving the problem will miraculously come into awareness.
* Because we do DDDM, doesn’t automatically translate into improved outcomes

…*thoughts*

* **No amount of increased effort** can compensate for limited knowledge and skill about how to solve problems that require special training. The only examples of accountability improving judgments and decisions requiring formal training are those where participants received training in the relevant decision rules for problem solving.

Moving beyond DDDM

Our **mission** is to support a program of research that seeks to understand how evidence-based programs developed to support students with persistent academic challenges can be tailored to assure the best possible outcome for every student. This tailoring acknowledges child, family & community needs and preferences – *local wisdom*.

Beyond DDDM

The emerging knowledge base will provide educators with options to deliver more efficient and effective instruction and services across diverse populations and settings. Advances in ***action research*** help extend the reach of evidence-based practices to underserved communities where they can reduce educational disparities.

The Notions of Equity



Rethinking Equity



http://culturalorganizing.org

**2016 MSS: Post High School Plans**

****

* Student-centered approaches show great promise
* Helping students set appropriate goals and supporting them to meet their goals show great promise
* Teacher commitment and teacher-powered schools show great promise
* The key ingredient in all effective youth-development efforts is ***developmental relationships***