7. Test Reliability

7.1. Precision and Reliability (Classical Methods)

Standard Error of Measurement (SEM) of the Test

Classical test theory is based on the following assumptions (Andrich & Luo, 2004):

- Each person *v* has a true score on the construct, usually denoted by the variable $T_{v.}$
- The best overall indicator of the person's true score is the sum of the scores on the items and is usually denoted by the variable $X_{v.}$
- This observed score will have an error for each person, usually denoted by $E_{v.}$
- These errors are not correlated with the true score.
- Across a population of people, the errors sum to 0 and they are normally distributed.

Based on these assumptions, useful indices are available within the framework of classical test theory (CTT) for estimating the precision of the raw test scores and the reliability of assessments. Within CTT, an observed test score is defined as an imprecise estimate of a student's true (and unobservable) proficiency level and is composed of two components. The first component is referred to as "true score" and is the portion of the observed score that is directly dependent on the student's proficiency level. The second is an error component (error) and is the portion of the score that is attributable to random error, that is, the portion of the score attributable to factors unrelated to the student's proficiency. Error for any student is normally distributed around that student's true score with a mean of zero and an arbitrary standard deviation. Suppose it were possible to give an exam to one student a large number of times without any practice effects. If we were to examine the resulting distribution of scores, we would find a normal distribution with a certain mean and a certain standard deviation about the mean. The mean of the resulting distribution is the student's true score according to the definition of error given above. For each student who responds to the exam, error is normally distributed with a mean of zero. However, the standard deviation of the error distribution is idiosyncratic to each student (though it tends to be larger toward the low and high ends of the exam for most tests). If we wanted to estimate what would likely be the standard deviation of this distribution of errors for any arbitrary examinee, the best estimate would be the mean of the standard deviations of the error distribution across all examinees. This quantity is called the standard error of measurement (SEM).

From the assumptions outlined and discussed above, the following mathematical formula can be derived:

$$X_v = T_v + E_v .$$

Therefore,

$$\sigma_x^2 = \sigma_t^2 + \sigma_e^2$$

where

σ_x^2 = the variance of the observed score in a population of persons,

σ_t^2 = the variance of their true score variance, and

σ_e^2 = the error variance.

The reliability coefficient of the test can be calculated by the following formula:

$$\rho_x = \frac{\sigma_t^2}{\sigma_x^2} = \frac{\sigma_x^2 - \sigma_e^2}{\sigma_x^2}.$$

Thus, the *SEM* is calculated by the following formula:

$$\sigma_e = \sigma_x \sqrt{1 - \rho_x}.$$

The SEM is commonly used in interpreting and reporting individual test scores and score differences on tests (Harvill, 1991). This equation, however, is only useful to estimate true score when the test reliability is reasonably high and the obtained score for the examinee is not an extreme deviation from the mean of the appropriate reference group. Consequently, when we use this equation, we should be careful with statements so that they do not imply greater precision than is actually involved (Harvill, 1991).

The SEM for each grade level of the test is provided in Chapter 9 in Table 9.1.1: Classical Descriptive Statistics for the 2011 Mod-MSA: Reading: Grades 3 though 8.

<u>Cronbach's Alpha (KR$_{20}$)</u>

Cronbach Alpha can be calculated by several methods. For dichotomously scored items, one of the best methods is the Kuder Richardson 20 (Crocker & Algina (1986), p.139) to estimate the internal consistency of items in the tests. Since the Mod-MSA: Reading tests include only SR items, the following formula was used to obtain the KR$_{20}$:

$$KR_{20} = \frac{k}{k-1}\left(1 - \frac{\sum pq}{\hat{\sigma}_x^2}\right)$$

KR$_{20}$ = Kuder Richardson 20

k = number of items on the test

pq = variance of item i, and

$\hat{\sigma}_x^2$ = total test variance

KR$_{20}$ is provided as reliability of the test in Table 9.1.1.

7.2. IRT Method in Measuring Precision of the Test

The information function (as discussed and provided in Section 9.4) is a function of proficiency and can be used to measure the precision of the test under IRT methods at a specified proficiency level. Conversely, the greater the information, the more precise will be the measurement of proficiency.

The inverse of the information function is the same as the conditional standard error of measurement (CSEM) discussed and provided in Section 9.4. The figures depicting CSEM provided in Section 9.4 show the standard errors of measurement at different proficiency levels of the examinees.

7.3. Decision Accuracy and Consistency at the Cut Scores

The accuracy and consistency analyses make use of the methods outlined and implemented in Livingston and Lewis (1995), Haertel (1996), and Young and Yoon (1998).

The *accuracy* of a decision is the extent to which it would agree with the decisions that would be made if each student could somehow be tested with all possible parallel forms of the assessments. The *consistency* of a decision is the extent to which it would agree with the decisions that would be made if the students had taken a different form of the examination, equal in difficulty and covering the same content as the form they actually took.

Students can be misclassified in one of two ways. Students who were below the proficiency cut score, but were classified (on the basis of the assessment) as being above a cut score, are considered to be *false positives*. Students who were above the proficiency cut score, but were classified as being below a cut score, are considered to be *false negatives*.

For the 2011 Mod-MSA: Reading, Tables 7.3.1 through 7.3.6 include:

- Performance level
- Accuracy classifications
- False positives
- False negatives
- Consistency classifications

The tables illustrate the general rule that decision consistency is less than decision accuracy.

Table 7.3.1. The 2011 Mod-MSA: Reading Decision Accuracy and Consistency Indices: Grade 3

Performance Cut	Accuracy	False Positive	False Negative	Consistency
B : PA	0.88	0.08	0.04	0.83
BP : A	0.94	0.05	0.01	0.91

Note. B:PA denotes the cut between Basic and Proficient, while BP:A denotes the cut between Proficient and Advanced. 2. These analyses are based on the statewide population after applying equating exclusion criteria

Table 7.3.2. The 2011 Mod-MSA: Reading Decision Accuracy and Consistency Indices: Grade 4

Performance Cut	Accuracy	False Positive	False Negative	Consistency
B : PA	0.85	0.09	0.05	0.80
BP : A	0.92	0.06	0.02	0.89

Note. B:PA denotes the cut between Basic and Proficient, while BP:A denotes the cut between Proficient and Advanced. These analyses are based on the statewide population after applying equating exclusion criteria

Table 7.3.3. The 2011 Mod-MSA: Reading Decision Accuracy and Consistency Indices: Grade 5

Performance Cut	Accuracy	False Positive	False Negative	Consistency
B : PA	0.87	0.07	0.07	0.81
BP : A	0.96	0.03	0.01	0.95

Note. B:PA denotes the cut between Basic and Proficient, while BP:A denotes the cut between Proficient and Advanced. These analyses are based on the statewide population after applying equating exclusion criteria

Table 7.3.4. The 2011 Mod-MSA: Reading Decision Accuracy and Consistency Indices: Grade 6

Performance Cut	Accuracy	False Positive	False Negative	Consistency
B : PA	0.85	0.09	0.06	0.79
BP : A	0.91	0.07	0.02	0.88

Note. B:PA denotes the cut between Basic and Proficient, while BP:A denotes the cut between Proficient and Advanced. These analyses are based on the statewide population after applying equating exclusion criteria

Table 7.3.5. The 2011 Mod-MSA: Reading Decision Accuracy and Consistency Indices: Grade 7

Performance Cut	Accuracy	False Positive	False Negative	Consistency
B : PA	0.86	0.09	0.06	0.80
BP : A	0.97	0.03	0.00	0.96

Note. B:PA denotes the cut between Basic and Proficient, while BP:A denotes the cut between Proficient and Advanced. These analyses are based on the statewide population after applying equating exclusion criteria

Table 7.3.6. The 2011 Mod-MSA: Reading Decision Accuracy and Consistency Indices: Grade 8

Performance Cut	Accuracy	False Positive	False Negative	Consistency
B : PA	0.84	0.10	0.06	0.78
BP : A	0.92	0.06	0.02	0.89

Note. B:PA denotes the cut between Basic and Proficient, while BP:A denotes the cut between Proficient and Advanced. These analyses are based on the statewide population after applying equating exclusion criteria