Item Response Theory, Reliability and Standard Error

Brent Culligan
Aoyama Gakuin Women's Junior College

When we give a test, it is usually because we have to make a decision and we want the results of the testing situation to help us make that decision. We have to interpret those results, and to make the case that our interpretations are valid for that situation. Validity, therefore, is an argument that we make about our assumptions, based on test scores. We must make the case that the instrument we use does, in fact, measure the psychological trait we hope to measure. Validity is, according to the Standards for Educational and Psychological Testing, "the most fundamental consideration in developing and evaluating tests" (cited in Hogan & Agnello, 2004).

One kind of support for the validity of the interpretation is that the test measures the psychological trait consistently. This is known as the reliability of the test. Reliability, i.e., a measure of the consistency of the application of an instrument to a particular population at a particular time, is a necessary condition for validity. A reliable test may or may not be valid, but an unreliable test can never be valid. This means that a test cannot be more valid than it is reliable, i.e., reliability is the upper limit of validity. It is important to remember that any instrument, i.e., the SLEP test or TOEFL, does not have "reliability." An instrument that demonstrates high reliability in one situation may show low reliability in another. Reliability resides in the interaction between a particular task and a particular population of test-takers.

While the reliability of a test is clearly important, it is probably one of the least understood concepts in testing. One of the purposes of the reliability coefficient of a test is to give us a standard index with which to evaluate the validity of a test. More importantly, the reliability coefficient provides us with a way to find the SEM, the *Standard Error of Measurement*. SEM allows practitioners to answer the question, "If I give this test to this student again, what score would she achieve?" In high stakes testing, this is a critical issue. A test taker gets 79. The cut-off is 80. Her life will take very

different paths based on your judgment. How confident are you of your test? Does she pass or fail?

In the first part of this paper, I will review how the reliability index, K-R20, and the *Standard Error of Measurement* are calculated under *Classical Test Theory*. I will then review the basic principles of *Item Response Theory*, and how the *Information Function* is used to obtain a *Standard Error of the Estimate*, a statistic similar to the SEM. I will conclude with an explanation of how this affects the scores reported by V-Check and how the scores can be interpreted.

Reliability and Standard Error of Measurement

One of the most commonly reported indices of reliability under *Classical Test Theory* is the Kuder-Richardson Formula 20, or K-R20.

Kuder-Richardson Formula 20

$$\text{K-R20} = \frac{k}{k-1}\left(1 - \frac{\sum pq}{s^2}\right) \tag{1}$$

where k is the number of items on the test

$\sum pq$ is the sum of the item variance

p is the total of correct responses divided by the number of examinees

q is the total of incorrect responses divided by the number examinees

s^2 is the test score variance

This formula is applied to dichotomously scored data. In theory, this reliability index ranges from +1.00 to –1.00. Reliability coefficients of over .80 are considered to be very good, and over .90 are excellent. To obtain the K-R20 index for a test, you must first find the sum of the variance for each item (pq) and the variance for the test scores. Remember that variance is a measure of the dispersion, or range, of the variable. Reliability, as measured by the K-R20 formula, is the result of these two factors, item variance, and test variance. The K-R20 reliability index is directly proportional to the variances of the test, i.e., if the sum of the item variance remains constant, as the test

variance increases, so too does the reliability. This is also why reliability by itself paints an incomplete picture, as we shall see in the next section.

Standard Error of Measurement

The *Standard Error of Measurement* attempts to answers the question, "If I give this test to this student again, what score would she achieve?" The SEM is calculated using the following formula:

$$SEM = s\sqrt{1-r} \tag{2}$$

where r is the reliability estimate of the test

s is the standard deviation of the test

We interpret the *Standard Error of Measurement* based on a normal distribution. That is to say, we would expect the score to be within 1 SEM 64% of the time, and to be within 2 SEM 98% of the time. When we speak of 95% confidence intervals, we are confident that the student would be within 1.96*SEM of the score 19 times out of 20. For example, a 100-item test with a K-R 20 of .90 (excellent!) and a standard deviation of 10 would have an SEM of $10\sqrt{1-.9}$ or 3.16. If a student had a score of 75, we would interpret this as follows. If the student took the test repeatedly, we would expect her scores to fall between 71.84 and 78.16 64% of the time. If we wanted to be 95% confident in the test scores, we would look at the interval of 75 ± 1.96(3.16). Now say we had another 100-item test with a K-R 20 of .60 (somewhat low) and a standard deviation of 5. The SEM would be $5\sqrt{1-.6}$ or 3.16 and we could interpret it exactly the same as the previous test.

Another way that we can interpret the SEM is that it shows us the error variation around the student's true score. In classical test theory, the observed score is composed of the true score plus error, and this error is normally distributed. Under this interpretation, the students' observed scores are with 1 SEM of the true scores 68% of the time.

Item Response Theory

Item response theory is a probabilistic model that attempts to explain the response of a person to an item (Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980). In its simplest form, item response theory posits that the probability of a random person *j* with ability θ_j answering a random item *i* with difficulty b_i correctly is conditioned upon the ability of the person and the difficulty of the item. In other words, if a person has a high ability in a particular field, he or she will probably get an easy item correct.

Conversely, if a person has a low ability and the item is difficult, he or she will probably get the item wrong. For example, we can expect someone with a large vocabulary to respond that they know easy words like 'smile' and 'beautiful' but we should not expect someone with a small vocabulary to know words like 'subsidy' or 'dissipate.' When we analyze item responses, we are trying to answer the question, "What is the probability of a person with a given ability responding correctly to an item with a given difficulty?" This can be expressed mathematically through a number of different formulae, but for this explanation I will focus on the One-parameter Logistic Model, also known as the Rasch (1960) Model, one of the most commonly reported in the literature.

One-parameter Logistic Model

Using the Rasch model, we can calculate the probability of an examinee answering an item correctly with the following formula:

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} \tag{3}$$

where $P_i(\theta)$ is the probability of a randomly chosen examinee with ability θ answering item *i* correctly.

e is the base of natural logarithms (2.718)

θ is the person ability measured in logits

b_i is the difficulty parameter of the item measured in logits

What is a logit?

A logit is a unit on a log-odds scale. The most important point to note is that IRT models are probabilistic. Because of this, we are interesting in finding out the odds of an event occurring, much like betting shops in Britain and Las Vegas. The definition of the odds of an event happening is the ratio of the probability of it occurring to the probability of it not occurring. For example, on a roulette wheel, there are 38 slots, so your probability of success is 1/38, and your probability of failure is 37/38. Your odds in favour of winning are 1 to 37.

$$\frac{1/38}{37/38} = \frac{1}{38} \bullet \frac{38}{37} = \frac{1}{37}$$

With IRT, the probability of an event occurring is the probability of correct response, or $P_i(\theta)$, and the probability of the event not occurring is $Q_i(\theta) = 1 - P_i(\theta)$, which is defined as the probability of a randomly chosen examinee with ability θ answering item *i* incorrectly (see Formula 4).

$$Q_i(\theta) = \frac{1}{1+e^{(\theta - b_i)}} \tag{4}$$

The odds of a correct response are $\frac{P_i(\,,\,)}{Q_i(\,,\,)}$.

$$\frac{P_i(\,,\,)}{Q_i(\,,\,)} = \frac{\frac{e^{(\theta - b_i)}}{1+e^{(\theta - b_i)}}}{\frac{1}{1+e^{(\theta - b_i)}}} = \frac{e^{(\theta - b_i)}}{1+e^{(\theta - b_i)}} * 1 + e^{(\theta - b_i)} = e^{(\theta - b_i)}$$

We can see that the odds in favour of a correct response is equal to $e^{(\theta - b)}$. Taking the natural log on both sides we get,

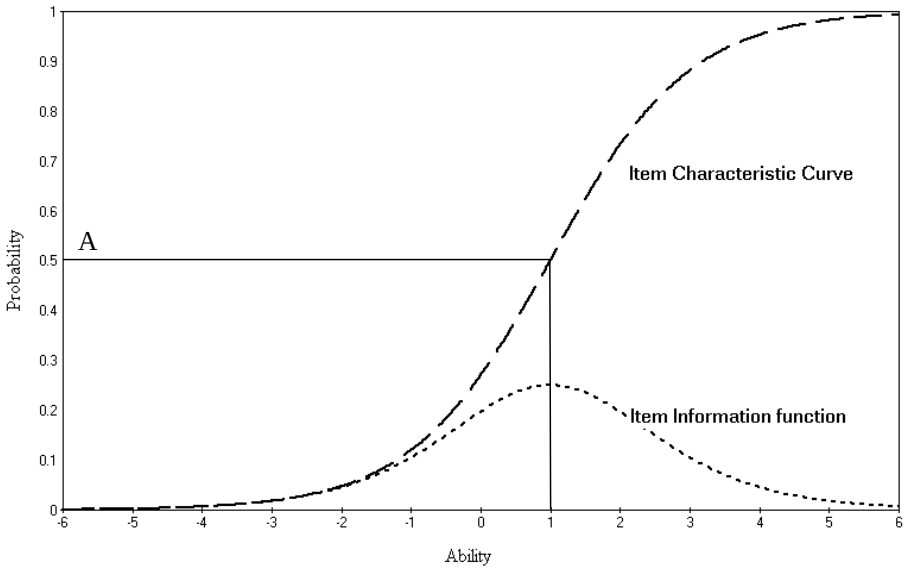
$$\ln \frac{P_i(\,,\,)}{Q_i(\,,\,)} = \ln e^{(\theta - b_i)} = \,, - b_i$$

The log of the odds is equal to $\theta - b_i$. , i.e. the difference between the ability of the student and the difficulty of the item measured in log-odd units, or logits (*L*). The higher the value of the estimate of ability, θ, the more ability the case, or person, has. The estimate of ability, θ, can range from $-\infty < \theta < \infty$. Likewise, the higher the value of the estimate of difficulty, *b*, the more difficult the item is. The estimate of item difficulty, *b*, can also range from $-\infty < b < \infty$. As the difference between θ and *b*

increases, the probability approaches zero, determine the probability of the answer being correct. Conversely as the difference decreases, the probability approaches 0. For example, at $\theta - b = 5$, the probability of a correct answer would be .9933, or at $\theta - b = -5$, the probability of a correct answer would be .0067.

We can plot the probability of a correct response to an item as a function of ability and get the *Item Characteristic Curve* (ICC). For example, in Figure 1, we can see the ICC of an item with difficulty 1.00, and this presents in visual form the information contained in the sixth column in Table 1. For most item analyses, a threshold, based on the probability of a correct response, is set by convention at .50. This threshold is significant in that it also establishes the location of the ICC, or where the ICC is located relative the horizontal ability axis (see Figure 1). Line A represents the threshold where persons of ability 1.00 have a .50 chance of getting the item correct. What this means is that the ability of a respondent is set when the respondent has a 50 percent probability of getting the answer correct. This occurs when the ability of the respondent matches the difficulty of the item.

Figure 1. Characteristic Curve and Information Function of the Item



Information Function

Also seen in Figure 1 is a graph of one of two information functions, the *Item Information Function*. Information functions are vital to the calculation of the *Standard Error of the Estimate* (SEE), an IRT statistic that can be interpreted in the same way as

the SEM of *Classical Test Theory*. The *Item Information Function* for a 1-parameter logistic model is defined as:

$$I_i(\theta) = P_i(\theta)Q_i(\theta) \tag{5}$$

This makes intuitive sense, if you look at the graph of the *Item Information Function* in Figure 1. A test item that everyone knows does not reveal very much information other than that everyone's ability is higher than the difficulty of the item. Likewise, the same is true for a difficult item that no one can answer correctly. Useful information is obtained only when some examinees know the item and when some do not.

The *Test Information I(θ)* is the sum of all the *Item Information* at θ.

$$I(\theta) = D^2 \sum_{i=1}^{n} I_i(\theta) \tag{6}$$

where D is a scaling factor of 1.70 to make the logistic probability distribution function similar to the normal probability distribution function

Finally, the SEE is equal to the reciprocal of the *Item Information Function*, i.e.,

$$SE(\theta) = \frac{1}{\sqrt{I(\theta)}} \tag{7}$$

With a computer adaptive test, the *Test Information* is calculated after the response to each item, and therefore can serve as a useful criterion for ending a test, once a sufficiently accurate score is obtained. For a test such as V-check, which had the *Test Information* set at 10, the SEE is about .32 logits. In order to obtain the comparable reliability index from a computer adaptive test, this SEE can be used. Thissen (2000) suggests the following formula,

$$r = 1 - SEE^2 \tag{7}$$

For V-check, this corresponds to a reliability coefficient of .90. V-check, however, does not report these logit scores, as they can be quite meaningless out of context. The logit scores are converted to the number of known words based upon a non-linear regression formula between the ability, as measured in logits, and the number of known words, as seen in Figure 2 (Thissen, 2000).

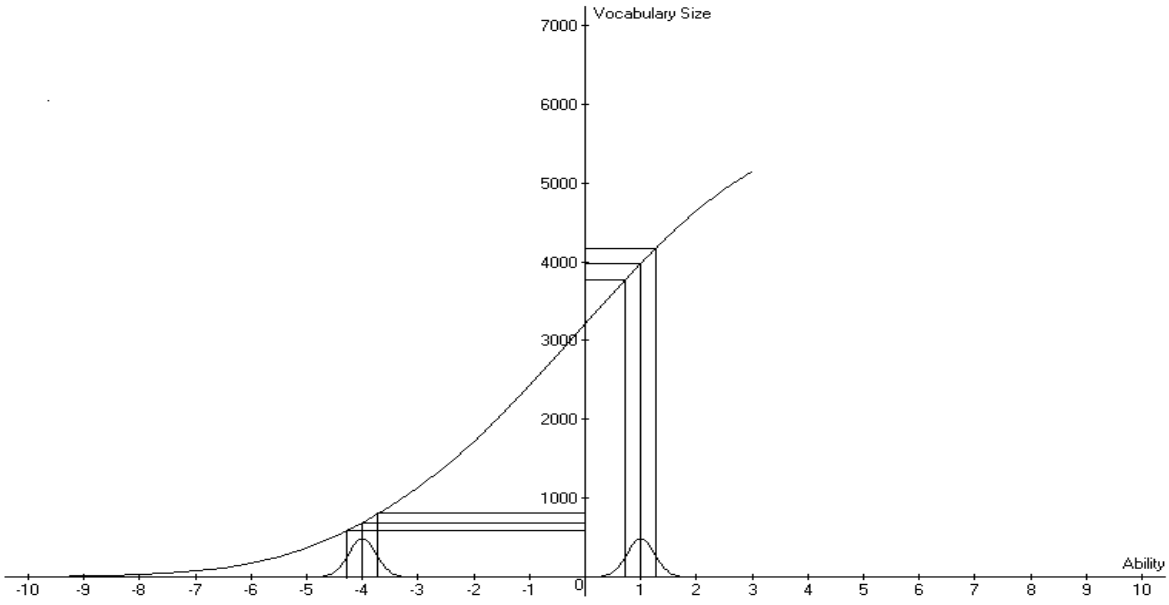


Figure 2. Non-linear Regression between Ability in Logits and Lexical Size

The following explanation pertains to both the SEE and the standard deviation (SD) of a group. In many way, the relationship between an individual's score and his or her SEM is the same as the mean of a group and the Standard Deviation. Both explain the expected value of scores using the normal distribution. As with the individual scores, the mean of the group is calculated in logits and converted to known words. However, the standard deviation (SD), like the standard error, is not reported because of the asymmetrical nature of its distribution. In Figure 2, the normal distribution of logit scores from a group is shown on the horizontal axis to illustrate how the mean and standard deviations, when measured in logits, are converted to the number of known words. As can be seen from the figure, the standard deviations are symmetrical around the mean. However, because the regression line is non-linear, this symmetry does not transfer to the number of known words. Table 1 presents a hypothetical class with a mean of .90 logits and a Standard Deviation of .50. In this class, 68% of the students' scores would fall between .40 and 1.40 logits. In terms of the number of known words, the mean of the ability would be converted into a mean of about 3,900 known words (rounded to the nearest tens). One standard deviation below the mean (.40) translates into a score of about 3,520 words, or 380 words below the mean. One SD above the mean is converted to 4,250 words or 350 words above the mean. Table 1 shows the gaps

between the mean and the converted number of words for 2 standard deviations above and below the mean. Because of this asymmetrical distribution, the use of standard deviations in statistical tests would be meaningless and would violate many of the assumptions of the statistical models. Given that the ability logit is a much sounder and more defensible measure than the number of words known, we plan to release a scaled score, known as a Lexxit, to interested researchers. Please contact us for details.

Table 1. Means, Standard Deviations and Number of Known Words

Interval	Ability	Known Words	SD in Words
-2SD	-0.1	3130	770
-1SD	0.4	3520	380
Mean	0.9	3900	
+1SD	1.4	4250	350
+2SD	1.9	4570	670

References

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

Hogan, T. P., & Agnello, J. (2004). An empirical study of reporting practices concerning measurement validity. Educational and Psychological Measurement, 64, 802-812.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Rasch, G. (1993). Probabilistic models for some intelligence and attainment tests, with a forward and afterword by B. D. Wright. Chicago: Mesa Press. (1960, Danish Institute for Education Research)

Thissen, D. (2000). Reliability and measurement precision. In H. Wainer (Ed), Computerized adaptive testing: A primer (2nd ed., pp. 159-183). Mahwah, NJ: Lawrence Erlbaum Associates.