SOME ISSUES RELATED TO THE USE OF CUT SCORES

Larry R. Nelson

Burapha University, Chonburi Thailand

Curtin University of Technology, Western Australia

Nelson, L.R. (2007). Some issues related to the use of cut scores. *Thai Journal of Educational Research and Measurement (ISSN 1685-6740): 5(1),* 1-16.

Cut scores are test scores used in mastery testing and in certification tests to identify those test takers who appear to have achieved at a predetermined level of proficiency. Cut scores are often used to differentiate test takers on a pass-fail basis. This paper takes up points raised by Berk (2000), discussing the use of reliability and error indices which are particularly appropriate for testing procedures which use cut scores. It is suggested that these indices are still not widely used. In an effort to promote their use and interpretation, results derived from instruments developed by testing centers in a variety of countries are presented.

Note, May 2013: One of the software systems mentioned herein, "Lertap", now has more support for mastery test users. Details here.

The *Standards for Educational and Psychological Testing*

The *Standards for Educational and Psychological Testing* have been around for decades. The latest edition appeared some eight years ago (AERA/APA/NCME, 1999).

Berk (2000) pointed out that there are now at least three standards which are directly concerned with the use of testing procedures involving cut scores:

Standard 2.14

Conditional standard errors of measurement should be reported at several score levels if constancy cannot be assumed. Where cutscores are specified for selection or classification, the standard errors of measurement should be reported in the vicinity of each cutscore.

Comment: Estimation of conditional standard errors is usually feasible even with the sample sizes that are typically used for reliability analyses. If it is assumed that the standard error is constant over a broad range of score levels, the rationale for this assumption should be presented. (p. 35)

Standard 2.15

When a test or combination of measures is used to make categorical decisions, estimates should be provided of the percentage of examinees who would be

classified in the same way on two applications of the procedure, using the same form or alternate forms of the instrument.

Comment: When a test or composite is used to make categorical decisions, such as pass-fail, the standard error measurement at or near the cutscore has important implications for the trustworthiness of these decisions. (p. 35)

Standard 14.15

Estimates of the reliability of test-based credentialing decisions should be provided.

Comment: The standards for decision reliability described in chapter 2 [2.14 and 2.15 above] are applicable to tests used for licensure and certification. Other types of reliability estimates and associated standard errors of measurement may also be useful, but the reliability of the decision of whether or not to certify is of primary importance. (p. 162)

Conditional standard errors of measurement

It has long been acknowledged that the standard error of measurement "varies across the range of examinee ability" (Qualls-Payne, 1992). The estimate of the standard error of measurement traditionally used in classical test theory is an average figure (see Feldt 1984 for a particularly cogent demonstration). When we have on hand a testing procedure with high-stakes outcomes, such as that encountered in the application of mastery and certification tests, it is important to recognize that the standard error of measurement in the vicinity of the cut score is likely to differ from the average. This concern is what is reflected in Standard 2.14.

There are several methods available for deriving conditional standard errors of measurement, that is, standard errors of measurement at different score levels. Thorndike (1951) suggested that one could take results from all test takers having a given observed score, compute an estimate of the measurement error variance for each respondent by partitioning the respondent's item responses into parallel half tests, computing the variance between the halves, and then averaging the result over all respondents at the given score level.

Lord (1955) suggested that conditional standard errors of measurement could be derived by application of the binomial error model. Qualls-Payne (1992) detailed how Keats (1957), Feldt (1984), and Jarjoura (1986) refined Lord's application of the binomial model. Lord himself has suggested another refinement (Lord, 1984). By and large these refinements involve a compound binomial model; in general they result in estimates of conditional measurement errors that are smaller than those obtained by applying the simple binomial model.

For a very readable description of the binomial error model, see Crocker and Algina (1986). Feldt's (1984) demonstration of the relationship between the binomial error model and classical test theory is also elucidating.

In terms of software capable of deriving conditional standard error of measurement estimates, readers could turn to Biddle's TVAP system (Biddle, 2003), or to the Lertap system (Nelson, 2000).

TVAP uses Thorndike's (1951) method. The TVAP manual provides a clear description of how the method is applied, and seems to suggest that Qualls-Payne

(1992) found Thorndike's method to be one which "generally produces very similar results" to methods which are based on revised binomial or compound binomial models.

Qualls-Payne (1992) used a set of criteria to evaluate six methods for computing conditional standard errors of measurement. Three were based on the compound binomial model, two on classical test theory (including Thorndike, as stated in the TVAP manual), and one on the three-parameter item response theory model. The most-recommended method was judged to be Feldt's (1984), one of the compound binomial methods. Thorndike's method was generally found to be the least preferred – its redeeming feature might be that it is easier to calculate than the other five approaches (my suggestion, not Qualls-Payne's).

Figure 1 displays a graph of selected results from Table 3 in Qualls-Payne (1992).

Figure 1: Qualls-Payne data with n = 359

There are seven lines in the graph, tracing standard error of measurement estimates over thirteen score intervals. Six lines correspond to the methods investigated by Qualls-Payne (1992). The dashed line is one I have added; it represents values calculated by applying the binomial error model:

$$S.E._{\cdot}^{2} = \frac{x(n-x)}{n-1}$$

The equation above provides an unbiased estimator of the squared standard error of measurement for a person with an observed score of *x* on a test with *n* items (see Lord, 1984).

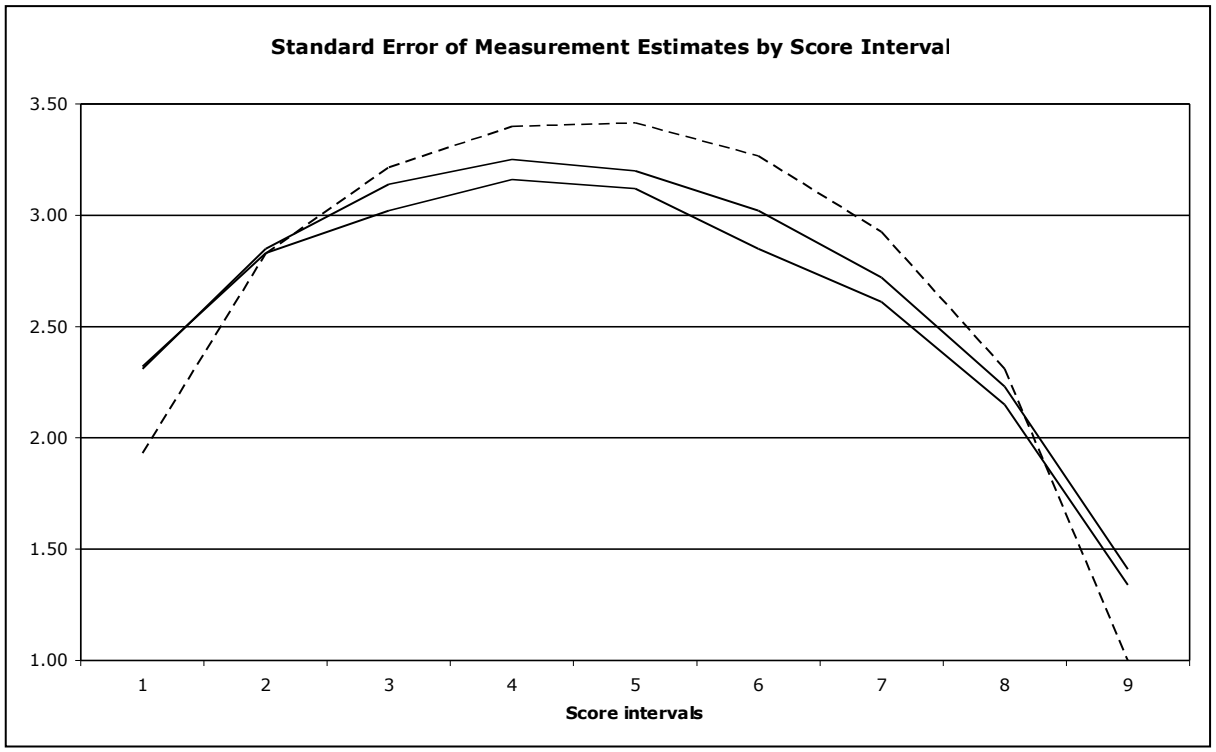
All but one of the lines in Figure 1 are rather smooth. The irregular line portrays the Thorndike estimates; it starts on the left with a dramatic dip, jumps around in a jagged manner, and eventually settles into something resembling a curve after we pass the

center of the score intervals. (Note: the Thorndike estimates go above a standard error of 3.00 for two of the score intervals; the other methods with estimates above 3.00 for some of the intervals are Feldt's, and the binomial error model. The line corresponding to estimates derived from the three-parameter item response model starts at the left at just under the 2.50 standard error line.)

A limitation of the Qualls-Payne study relates to sample size. Less than 400 students were tested. The sample sizes for several of the score intervals were meager; the first three intervals had n's of less than 20, while the next four intervals were each populated by less than 40 test takers.

Thorndike's procedure returned a poor "curve" in Figure 1, but there is evidence which shows that it can produce stable estimates given adequate sample size. For example, Figure 2 displays a graph of results from Table 1 of Feldt (1984).

Figure 2: Feldt data with n = 15,546



Feldt's (1984) paper involved comparing results from a compound binomial error model with those empirically obtained by application of Thorndike's (1951) method to more than 15,000 test takers. Feldt used nine test score intervals in his study.

In the graph seen in Figure 2, the two solid lines are from the Reading of Literary Materials test data found in Table 1 of Feldt (1984). Looking immediately above score interval "9", the highest line corresponds to Feldt's compound binomial (at a standard error of 1.41); the middle line corresponds to Thorndike's method (standard error 1.34); the dashed line is one I have imposed on the graph using the binomial error model, assuming that the total number of items on the test was 46.

Feldt (1984) stated that the similarity between the compound binomial and Thorndike estimates of the standard error "…represents strong support for the compound binomial error model….". The two lines are close over all score intervals; in fact, on the left, they merge for the first two intervals. Thorndike's method can obviously return a

relatively smooth curve when sample size is large. But, when given the sample sizes found in "everyday" testing, with n's below a thousand, we might assume that Thorndike's method could well result in coarse estimates for scores not close to the mean, as seen previously in Figure 1. This is an issue of obvious relevance to TVAP users.

I have superimposed the dashed line in Figure 2 to enable comparisons with the binomial error model. The limitations of this model have been cited by Lord, Feldt, and others – in the main, the model produces standard error estimates which are often too high, particularly in central score intervals.

The binomial error model assumes tests are comprised of random samples of items from an infinite (or very large) pool. Lord (1984) pointed out that the estimates obtained from the model are "too large" when test forms are created by matching items on the basis of common statistics, such as difficulty and discrimination. When this is the case, Lord (1984) suggested another approach, "Method IV", which adjusts the binomial model by assuming a compound binomial distribution of observed scores, and then determining the value of a constant, "K", which will make the index of reliability (the correlation between observed scores and true scores) equal to the square root of KR-20[1]; once derived, the constant is applied as a multiplier on the binomial error estimate.

Feldt (1984) also reported that the binomial error model produced figures which had been "criticized as being too large". When tests are constructed by stratified item sampling, instead of simple random sampling, Feldt suggested that application of the compound binomial model is justified, and he derived an empirical method for estimating standard errors based on the compound model.

As may be seen in Figure 2, the binomial error model estimates traced by the dashed line were considerably higher except at extreme score values. However, such was not the case with the Qualls-Payne data graphed in Figure 1; in Figure 1, the dashed line tracing the binomial error model estimates was not noticeably higher than the other lines. This may have resulted from the inadequate interval sample sizes already cited, and to a smaller number of test items (39 in the Qualls-Payne study, compared to 46 in Feldt's).

I have mentioned that the TVAP software package uses Thorndike's procedure to derive conditional standard errors of measurement. The Lertap item analysis system (Nelson, 2000) also computes conditional standard error of measurement estimates, but not with Thorndike. Instead, Lertap uses two other methods: the simple binomial error model, and the compound binomial model seen as "Method IV" in Lord (1984). I elected to use Lord's Method IV as it is relatively easy to calculate, and has recently been used by others (see, for example, the manual for the *Test of English for International Communication*, an instrument developed by an affiliate of the Educational Testing Service (www.toeic.com)).

Figure 3 displays a graph based on a standard Excel[2] chart as produced by Lertap for one of the data sets mentioned later in the paper.

The dashed curve in Figure 3 corresponds to standard error estimates made with the binomial error model. In this case there were 70 test items. The solid curve represents estimates made with Lord's (1984) Method IV. The straight line is the conventional standard error of measurement common to classical test theory, computed from the equation

[1] Kuder-Richardson formula 20.
[2] Microsoft Excel, a spreadsheet program.

$$SEM = [s_x^2(1 - r_{xx})]^{1/2}$$

where s_x^2 is the variance of observed scores, and r_{xx} the reliability of the measurement process.

Lertap computes coefficient alpha as its reliability estimate; alpha is equivalent to KR-20 when items are scored on a number-right basis. In this case, alpha was found to be 0.83, with SEM at 3.32.

The SEM for the test scores plotted earlier in Figure 2 was 2.50, with KR-20 at 0.84 (Qualls-Payne, 1992, Table 2). Both Figure 2 and Figure 3 indicate that standard error estimates computed at specific scores or score intervals were often greater than respective conventional SEM values.

When this discussion is placed within the context of cut scores, it is readily seen that the computation of conditional standard errors of measurement has practical relevance.

Suppose, for example, that a passing point of 70% had been set for the scores plotted in Figure 3. There were 70 items on the test, making the cut score 49. The two conditional standard error of measurement figures for a score of 49 were 3.47 and 3.86; the first of these figures derives from application of Lord's (1984) Method IV, and is a value appropriate in the usual case, when test items have been selected on the basis of specific criteria, such as subtopic, difficulty, or discrimination. The second figure, 3.86, is appropriate when items have simply been randomly sampled from an item pool, with no attempt to select them using a criterion.

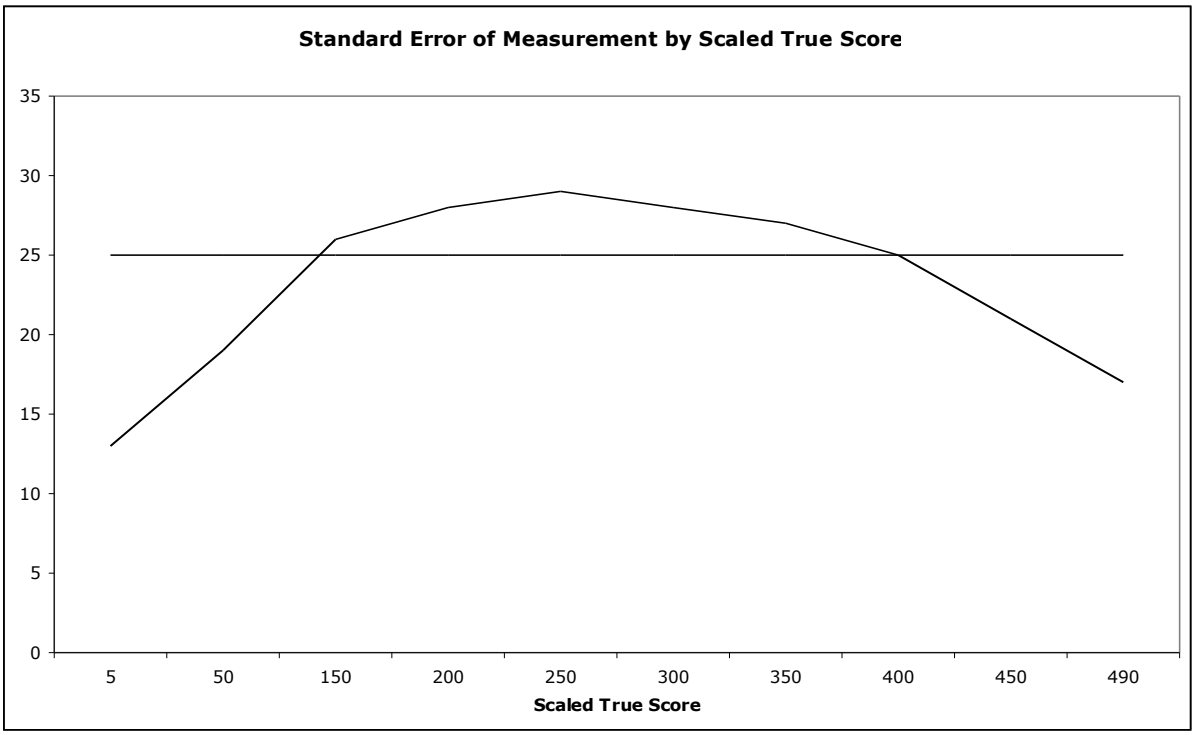
Figure 3: Lertap standard errors graph, n = 609

Both figures are greater than 3.32, the SEM value commonly computed in classical test theory. Were we to form 95% confidence intervals around an observed score of 49,

adding and subtracting 1.96 standard errors, we would obtain an interval of (42.5 to 55.5) from SEM, (42.2 to 55.8) from Lord's Method IV, and (41.4 to 56.6) from the binomial error model. The latter two intervals would become wider were the passing point lowered to a cut score of 43 (60%) – corresponding standard errors at this point are 3.69 and 4.10. Drop the cut score to 35 (50%), and the conditional standard errors increase to 3.79 and 4.21.

For comparison purposes I have graphed data from the Listening Comprehension TOEIC test (www.toeic.com) in Figure 4.

Figure 4: Data from TOEIC Listening Comprehension scale



The flat line in Figure 4 corresponds to the SEM derived from classical test theory, while the other line traces conditional standard error values computed at ten points using Lord's (1984) Method IV (using data as reported in the TOEIC Technical Manual (www.toeic.com)).

Again the relevance of using conditional standard errors with cut scores is implicit in Figure 4. Measurement error varies over the score range. We will have less confidence when using cut scores placed in the center of the possible score range. Use of the conventional SEM from classical test theory seems particularly inappropriate with very low or very high scores, where the SEM substantially over-represents measurement error.

Classification consistency

Standard 2.14, regarding the use of conditional standard errors of measurement at cut scores, was introduced in the 1999 version of the *Standards for Educational and Psychological Testing*. Standard 2.15 was also introduced in 1999; it focuses on using a measure of classification consistency at the cut score. In the words of Standard 2.15, we are to employ an estimate of "the percentage of examinees who would be classified

in the same way on two applications of the procedure, using the same form or alternate forms of the instrument".

Berk (2000) suggested that two indices of classification consistency would be candidates for the tool mandated by Standard 2.15: the "p_0 index", and "kappa".

The p_0 index is simple. We have a group of test takers sit the same test twice, or sit two parallel forms of a test. We then calculate the proportion of those above the cut score on both occasions, and add to it the proportion of those below the cut score on both occasions. This gives us p_0. If, for example p_0 = .8, the interpretation is that 80% of the test takers have been placed in the same group on both test sittings – 80% of the test takers have been consistently classified. For references, see Swaminathan, Hambleton and Algina (1974), Subkoviak (1976, 1988), and what is called $\hat{P}$ in Chapter 9 of Crocker and Algina (1986).

It is the case that this type of consistency estimate will be inflated by chance, a fact which has led to the use of what is referred to as "kappa", or as the "kappa coefficient", or, in Crocker and Algina (1986), as "Cohen's Kappa". Briefly, kappa adjusts p_0 by taking out the proportion of correct classifications which would be expected by chance alone.

In referring to these two measures, p_0 and kappa, Berk (2000) wrote that "*… p_0 is an unbiased estimate of decision consistency that is simple to compute, and explain; kappa is a biased estimate with a long list of limitations and statistical conditions which complicate its interpretation….*". An earlier article by Berk (1980) had a more extensive comparison of p_0 and kappa.

An obvious limitation of p_0 is that it requires two administrations of a test. Berk (1980) reviewed methods for estimating p_0 given a single test administration; of the methods addressed, he recommended Huynh's (1976).

Huynh's method is based on fitting a two-parameter beta binomial model to observed scores. Hanson and Brennan (1990) suggested that the two-parameter beta binomial does not always give a sufficient fit to observed scores, finding that a four-parameter beta binomial model was at times superior, depending on the shape of the distribution of observed scores. Brennan (2004) released a computer program, "BB-Class", which may be used to compute p_0 using both the two-parameter and four-parameter models.

A more general problem with Huynh's method is that it is not easy to compute; however, Peng and Subkoviak (1980) developed and tested a simplified approach to Huynh's method, finding it to provide sufficiently adequate estimates.

Lertap 5 has used the Peng-Subkoviak estimate of p_0 since it was initially published in 2000. In an effort to validate the use of the Peng-Subkoviak estimate, I compared Lertap's results with those from BB-Class. Results are shown in Table 1.

Table 1: Consistency indices for 12 data sets

Data	N	Nits	pass	skew	kurt	P-S	2-P	4-P
D2p	11190	25	3%	0.71	1.20	0.96	0.96	0.96
D2m	11190	25	15%	0.98	1.05	0.85	0.86	0.90
C2	3142	60	47%	-0.16	-0.50	0.81	0.80	0.81
B5	59	60	54%	-0.15	-0.92	0.78	0.76	0.85
B7	57	50	54%	-1.20	1.45	0.82	0.81	0.80
D1	1494	20	54%	0.10	-0.49	0.76	0.75	0.75
C1	603	70	58%	-0.26	-0.44	0.82	0.80	0.81
B2	649	60	64%	-0.60	-0.04	0.84	0.83	0.84
B6	57	60	70%	-0.41	-0.75	0.82	0.80	0.85
B3	128	40	71%	-0.68	0.43	0.83	0.83	0.83
B4	127	40	75%	-0.95	1.16	0.82	0.82	0.84
B1	267	150	83%	-1.10	1.51	0.92	0.92	0.93

Table 1's results derive from the application of two computer programs, Lertap 5 (Nelson, 2000), and BB-Class (Brennan, 2004), to twelve data sets.

The data sets are coded by source, with results presented in order of increasing pass rate. The two "D" data sets are from screening exams used by a large engineering school in Latin America. These exams were short, having 25 items each (Nits). They were designed to be difficult. With passing scores set at 13 for each, only 3% of the 11,190 first-year applicants passed the physics exam, while 15% passed the mathematics test.

The "B" exams are from licensing tests used in the southeast of the United States; in all cases, passing a B exam required a minimum score of 70%. The two "C" exams are from certification tests used in Australasia; these tests also had their cut score set at 70%.

The "skew" and "kurt" columns represent skewness and kurtosis, respectively.

The P-S, 2-P, and 4-P columns have values for three p_0 estimates: Peng-Subkoviak from Lertap, the two-parameter beta binomial from BB-Class, and the four-parameter beta binomial from BB-Class.

The P-S and 2-P values align over all of the data sets. However, the 4-P values differ in some cases, particularly in data sets D2m, B5, and B6.

Hanson and Brennan (1990) compared 2-P and 4-P estimates for eight ACT test sittings. They concluded that p_0 was overestimated by the 2-P model whenever the passing rate exceeded 50%, and underestimated by the 2-P model whenever the cut score resulted in a passing rate below 50%.

The results in Table 1 do not confirm the pattern reported by Hanson and Brennan (1990). I have looked carefully at the data, using the wealth of statistics provided in BB-Class output (including goodness-of-fit measures), and examining a variety of Q-Q plots

by applying SPSS[3]. I am unable to find a general pattern, except for this: if the score distribution has a "tail which wags", the 4-P model consistently provides a better fit, resulting in higher p_0 estimates. By "tail which wags" I mean a distribution which is not only skewed, but having a prominent cluster of cases at the skewed end. The B5 data set had such a tail, and also a bimodal tendency.

Hanson and Brennan (1990) reported that the improved fit sometimes provided by the 4-P model impacted kappa more than p_0. I also found this to be the case. For example, in the D2p data set, Lertap and the 2-P model returned kappa values of 0.17 and 0.16, respectively, when the 4-P model had kappa at 0.30. I have not elaborated on the kappa calculations here as, like Berk (2000), I consider p_0 to be a preferred statistic due to its ease of interpretation.

Reliability of credentialing decisions

Standard 14.15 involves estimates of the reliability of test-based credentialing decisions. The comment attached to the Standard states that "*…the reliability of the decision of whether or not to certify is of primary importance….*".

Assuming that the matter of awarding a credential is tantamount to testing with the use of a cut score, then, as Berk (2000) implies, the procedures discussed above are all relevant. Conditional standard errors of measurement, p_0, and kappa all bear on this Standard.

But there are other procedures of relevance here, and Berk (2000) mentions one of the most salient: a method based on a generalizability-theory (or decision-theory) approach. Specifically, Berk suggests that Brennan's (1992) $\Phi(\lambda)$ index is relevant to Standard 14.15.

Crocker and Algina (1986) would obviously agree, even though the Standard had not been published at the time of their text. In their Chapter 9, Crocker and Algina pointed out that p_0 and kappa "*treat all inconsistent classifications as equally serious*" (p. 203). They suggested that two other indices would be appropriate "*when the test developer wants to reflect the magnitude of misclassification in judging the reliability of decisions*" (p. 203): Livingston's coefficient, and Brennan and Kane's (1977) M(C) index. Of these two, Crocker and Algina wrote that M(C) would be preferred as it accounts for more sources of potential error. M(C) is referred to by its developers as "*an index of dependability for mastery tests*" (Brennan and Kane, 1977, p. 279).

Brennan's (1992) $\Phi(\lambda)$ index referred to by Berk, and Brennan and Kane's (1977) M(C) index referred to by Crocker and Algina, are the same. Shavelson and Webb (1991) made reference to the Brennan and Kane (1977) article, unambiguously referring to the index of dependability as Φ. Dimitrov (2003) wrote "*Brennan and Kane (1977) introduced a dependability index,* $\Phi(\lambda)$ *....*" Both of these references are incorrect. Brennan and Kane's (1977) article never used the Greek letter Φ for the index of dependability; it was consistently denoted as M(C). It is apparent that the preferred notation for the Brennan-Kane dependability index changed at some point, from M(C) to $\Phi(\lambda)$, probably in Brennan (1980).

[3] Statistical Package for the Social Sciences, www.spss.com.

Such notational discrepancies are of course not important in and of themselves. I have mentioned them in an effort to clarify this area of the literature, and to assist users of the Lertap 5 software system; the Lertap manual uses the term M(C), not $\Phi(\lambda)$, when it refers to the index of dependability.

The magnitude of the index of dependability depends on the cut score. The C in M(C) and the λ in $\Phi(\lambda)$ refer to the value used for the cut score, expressed as a proportion. For example, M(.50) and $\Phi(.50)$ would refer to the value of the dependability index when the cut score has been set to 50% of the maximum possible test score.

Figure 5 graphically portrays the relationship between $\Phi(\lambda)$ and cut score. The straight line in Figure 5 corresponds to the value of coefficient alpha for the data whose standard errors of measurement are plotted in Figure 3 above; in this case α = 0.83. The other line in Figure 5 traces the value of $\Phi(\lambda)$ for various values of λ, the cut score. (Note that the cut score values seen along the abscissa of Figure 5 are in fact proportions – the decimal point has been omitted.)

The average test score for the data underlying the graphs in Figure 3 and Figure 5, as a proportion of the maximum possible score, was .71, corresponding to a raw test score of 49.85.

Figure 5: Dependability index, $\Phi(\lambda)$, by cut score value

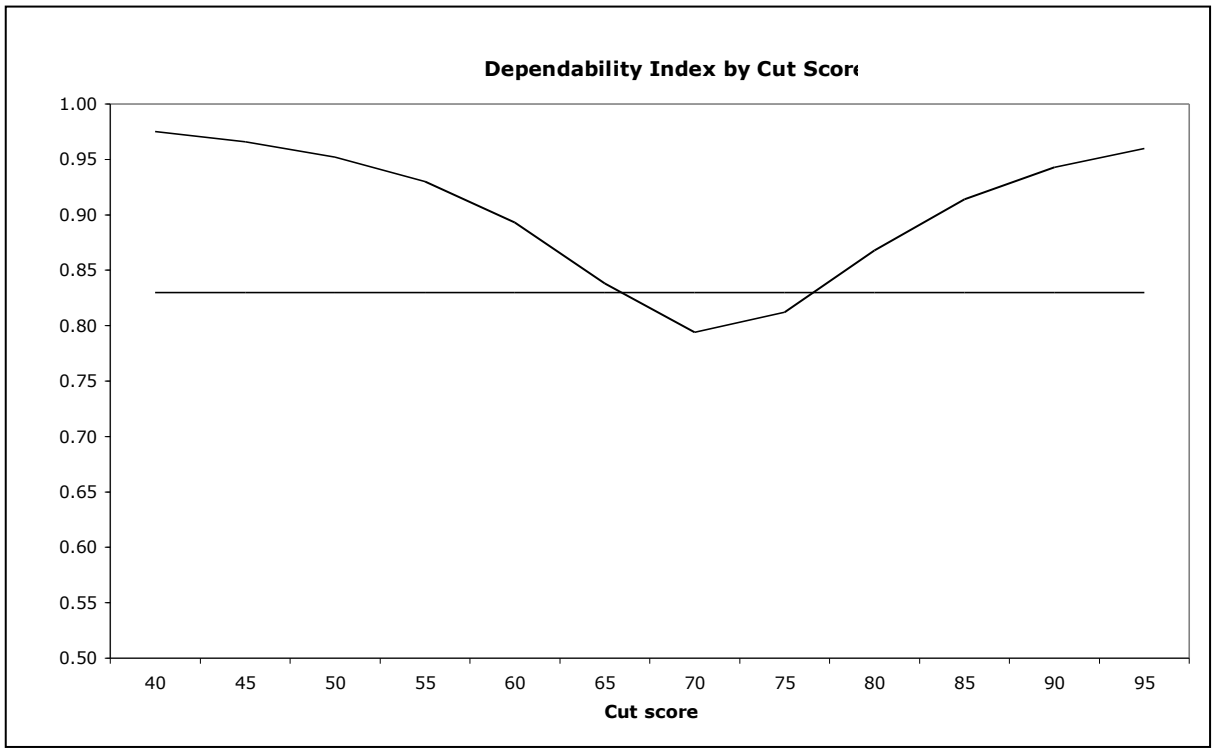


Figure 5 is similar to the graph of $\Phi(\lambda)$ seen in Brennan and Kane (1977). $\Phi(\lambda)$ dips to its lowest possible value near the average test score; at this point the value of the dependability index is lower than coefficient alpha. However, as we move away from the average, $\Phi(\lambda)$ increases – we could say that our measurement process becomes more dependable; the accuracy of classification decisions based on a cut score will be greater as the cut score moves away from the average test score.

When the $\Phi(\lambda)$ notation has been used for the dependability index, Φ, by itself, refers to the value of the dependability index when the cut score has been set equal to the test mean. When this is the case the literature will sometimes refer to "index Φ" (see, for example, Dimitrov, 2003). Index Φ is the lowest value that the dependability index may obtain for any given testing situation.

Both p_0 and $\Phi(\lambda)$ are measures of classification consistency. They are at times referred to as reliability coefficients for mastery tests (see, for example, Brennan and Kane, 1977, p. 286); Berk (1980) suggested they be referred to as "agreement indices". They use the same zero to one scale, with one being best. But they are not identical, as Crocker and Algina (1986) and others have highlighted. Brennan and Kane (1977) have a good description of the differences: M(C) assumes a squared-error loss function, while p_0 is based on a threshold loss function. They wrote (p. 287): "*A threshold loss function is appropriate when there is a sharp cutoff, and all classifications are, at least approximately, equal in their impact. A squared-error loss function is likely to be more appropriate when either of these assumptions is violated.*" To follow an example seen in Brennan and Kane (1977), suppose we are using a cut score of 90, and have two students, one with a true score of 20, another with a true score of 89. Suppose that our measurement process has erroneously misclassified both students. Statistics based on a threshold-loss function, such as p_0, make no distinction as to the gravity of such errors; to them the impact of misclassifying the true score of 20 is the same as the impact of misclassifying the true score of 89 – both errors are equally serious. Not so with a statistic based on a squared-loss function, such as $\Phi(\lambda)$, which is, in the words of Brennan and Kane (1977, p. 287) "sensitive to the magnitude of errors".

Crocker and Algina (1986) used the term "degrees of consistency" (p. 212), suggesting that p_0 is "perhaps the simplest measure of consistency of mastery decisions" (p. 212), but noting that it does not reflect the *degree* of consistency; should we be interested in degree, then an index such as $\Phi(\lambda)$ is appropriate. Crocker and Algina's text has a practical example of how these two statistics are affected by the magnitude, or degree, of classification errors (Crocker and Algina, 1986, Table 9.6, p. 205).

As a matter of interest, when I had Lertap derive the values of $\Phi(\lambda)$ plotted in Figure 5, I took note of corresponding p_0 values. Data are displayed in Table 2.

Table 2: Two consistency indices by cut score

Cut score	$\Phi(\lambda)$	p_0
.40	0.975	0.996
.45	0.966	0.989
.50	0.952	0.971
.55	0.930	0.938
.60	0.893	0.892
.65	0.838	0.844
.70	0.794	0.815
.75	0.812	0.819
.80	0.868	0.854
.85	0.914	0.903
.90	0.943	0.947
.95	0.960	0.976

Figure 6: Graph of $\Phi(\lambda)$ and p_0 values seen in Table 2

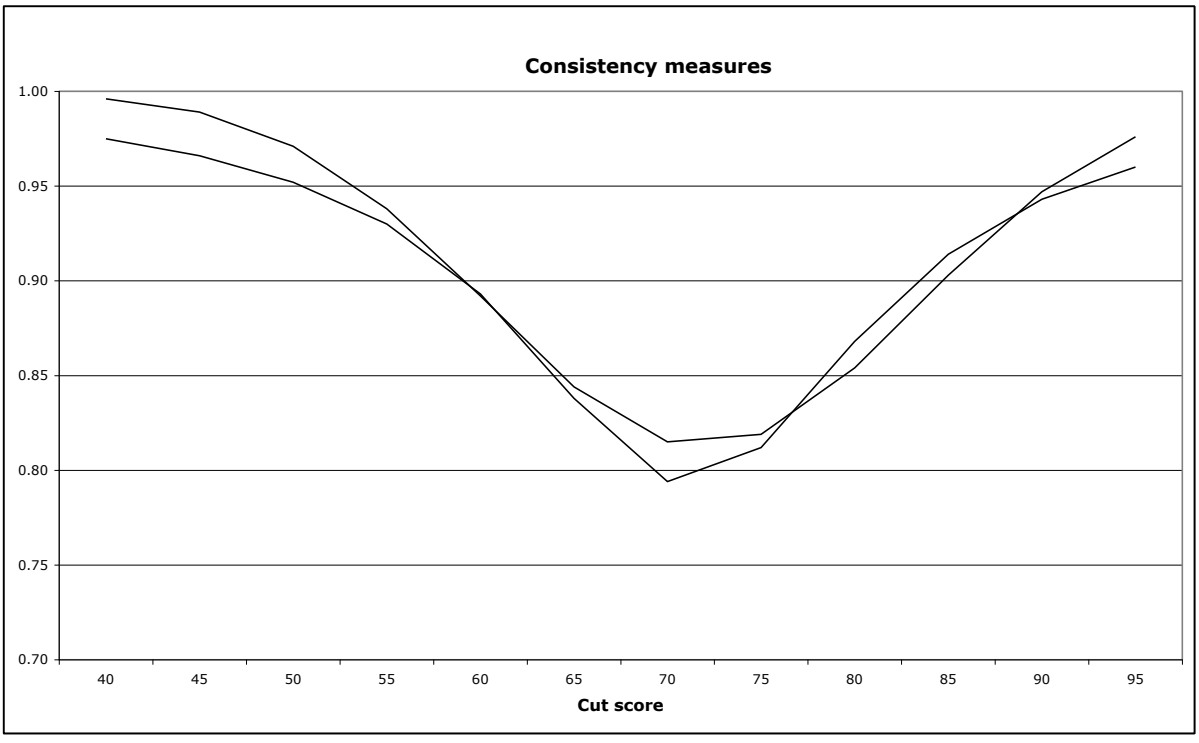


Figure 6 plots the values found in Table 2. (Note that the p_0 values in Table 2 were computed using the Peng-Subkoviak estimation method.)

As indicated in Figure 6, differences in the two consistency measures, p_0 and $\Phi(\lambda)$, were slight in this case. However, it cannot necessarily be concluded that this is a typical situation. p_0 is an indicator of the proportion of consistent classifications; $\Phi(\lambda)$ reflects not only this proportion, but also the magnitude of any misclassifications. It should not be concluded that these measures will always be as similar in value as those seen in this case.

Of general interest is the pattern seen in Figure 6: both measures dip close to the mean of the observed scores, and climb as the cut score departs from the mean – *this* is a general pattern, and a typical situation.

Putting things together

To re-cap, the latest edition of the *Standards for Educational and Psychological Testing* (AERA/APA/NCME, 1999) has at least three standards which apply to the use of cut scores. We should compute the standard error of measurement at the cut score, employ an estimate of the percentage of examinees who have been consistently classified on both sides of the cut, and, when appropriate, use a measure of the reliability of test-based credentialing decisions.

Methods for accomplishing these tasks have been with us for decades, but Berk (2000) reminded us of the slow rate at which these methods have been adopted by practitioners; with reference to the 1999 *Standards for Educational and Psychological Testing*, Berk wrote that he hoped "…Standards 2.15 and 14.15 will have a greater impact on practice…." (2000, p. 187).

At times I receive questions from Lertap 5 users related to mastery testing and cut scores. A general theme in these questions will have to do with obtaining a value for coefficient alpha on the low side, that is, say, less than .80. A related theme will generally relate to having test items which turned out to be quite easy, with the majority of item difficulty values at or above .70, accompanied by low item discrimination values (less than .20).

We know that tests with a majority of easy items will often return low alpha and item discrimination values when put into action. This does not mean that the tests are necessarily deficient. As Berk (2000) pointed out, it may be partly a matter of using the wrong statistics – if our interest is with assessing mastery, as in licensing and certification efforts, then we should deploy the methods recommended in the *Standards*.

I point Lertap users to Berk's (2000) document, and to those sections of the Lertap manual (Nelson, 2000) which are relevant to mastery testing. However, we could very well do with some baseline data. Berk and Nelson may be useful, but, in practice, what have others found when they've applied the "new" methods promulgated in the most recent *Standards*?

Table 3 is a partial answer. It provides more data for the twelve tests first seen in Table 1 above.

Table 3: Test statistics for twelve data sets

Data	avg.	cut	pass	alpha	SEM	CSEM	p_0	$\Phi(\lambda)$
D2p	.25	.50	3%	0.51	2.04	2.44	0.96	0.91
D2m	.33	.50	15%	0.77	2.08	2.37	0.85	0.86
C2	.68	.70	47%	0.83	3.21	3.28	0.81	0.80
B5	.71	.70	54%	0.77	3.04	3.14	0.78	0.70
B7	.68	.70	54%	0.84	2.94	3.04	0.82	0.82
D1	.50	.50	54%	0.73	1.96	2.12	0.76	0.68
C1	.71	.70	58%	0.83	3.32	3.47	0.82	0.79
B2	.73	.70	64%	0.86	3.06	3.31	0.84	0.84
B6	.74	.70	70%	0.79	2.87	3.10	0.82	0.76
B3	.76	.70	71%	0.80	2.44	2.78	0.83	0.82
B4	.76	.70	75%	0.77	2.47	2.79	0.82	0.80
B1	.79	.70	83%	0.92	4.44	5.21	0.92	0.95

The "avg." column in Table 3 indicates the average proportion-correct score for each test, while "cut" refers to the minimum proportion correct score required in order to pass the test. "Alpha" refers to the value of coefficient alpha on the occasions when each test was used, while "SEM" corresponds to the respective standard error of measurement.

The statistics recommended in Berk (2000) are found in the CSEM, p_0, and $\Phi(\lambda)$ columns; "CSEM" corresponds to the conditional standard error of measurement at the cut score, p_0 is the classification consistency estimate using the Peng-Subkoviak method, and $\Phi(\lambda)$ is the value of the dependability index at the cut score.

It is interesting to note that all of the CSEM figures exceed the corresponding SEM values. This is the expected case when the cut score is close to the center of the possible score distribution, that is, close to .50. Note that these results run contrary to the TVAP manual, where it is suggested that "*The CSEM is typically smaller than the SEM, since it is only taking the scores around the critical score into consideration*" (Biddle, 2003, p. 13). This suggestion may not be generally correct; we can expect the CSEM to be smaller at the extreme ends of the possible score range, but not necessarily at the critical (cut) score unless it is located close to one of the ends.

Another interesting point is that we can have a low alpha figure, yet obtain high p_0 and $\Phi(\lambda)$ values – see, for example, data set D2P in Table 3. Of note is that this data set's cut score of .50 resulted in a pass rate of just 3%. We expect p_0 and $\Phi(\lambda)$ to take on their highest values in the tails of score distributions; the same outcome would likely be observed with the pass rate at 97%.

In general, in nine of the Table 3 data sets, p_0 and $\Phi(\lambda)$, are close; the exceptions are data sets B5, D1, and B6.

The p_0 values seen in Table 3 indicate that, by and large, the application of the twelve tests has returned classification consistencies of about 80% or greater. This may

at first seem comfortable – the implication is that 80% of our pass / fail classifications would seem to be correct, or at least consistent; were we to measure the same people with the same instrument, we would expect 80% of them to end up in the same category, pass or fail.

However, one in five, or 20%, would not. Is this not a sobering corollary? The data sets I have used in this paper correspond to tests which have been professionally developed, and actively used as part of the process of passing or failing thousands of students. Yet we have solid evidence here to suggest, in a most practical metric, a consequence of applying most of the tests: a 20% inconsistency rate. Based on the evidence to hand, were we able to re-test the students, on many of these tests we might well expect about one in every five students to end up in a different pass / fail category.

What have others found when p_0 has been used?

Hanson and Brennan (1990) reported data from the use of two ACT tests, mathematics, with 40 items, and natural science, with 52 items. They found that, with the cut score set so that the pass rate was around 50%, p_0 was about .84 in the case of mathematics, and about .83 for the administration of the natural science test.

Subkoviak (1988), referring to p_0 as the "agreement coefficient", posed the question: "*What value of the agreement coefficient is satisfactory?*". His answer was that it "*…depends on the seriousness of the decisions being made with the test, and what can realistically be expected of a test in a given situation ….*" (p. 51). He went on to suggest this guideline: "*Tests used to make serious decisions should be sufficiently long to guarantee an agreement coefficient exceeding .85*" (p. 52).

Subkoviak couched his guideline in terms of test length as longer tests, when administered, generally return a higher reliability figure (such as alpha); Subkoviak's 1988 article drew a careful link between reliability and the value of the "agreement coefficient", and his message is certainly relevant: the longer a test, the more justification we have to expect an adequate p_0 value when the test is put to use. (One factor behind the high p_0 value for data set B1 in Table 3 is test length – test B1 consisted of 150 items.)

We also expect p_0 to rise as we move the cut score away from the average score, and (generally) as the pass rate departs from .50.

We could pose similar questions about the dependability index, $\Phi(\lambda)$. What have others reported? A literature research in this area has unfortunately not returned a great number of hits. It is apparent that the dependability index has been put to use, but actual values of the index have not quickly come to the fore in my searches to date.

I relate outcomes from just one study: Brown and Hudson (2002) reported $\Phi(\lambda)$ figures for the application of four reading tests, each with two forms. With cut scores at .90 (90%), reported $\Phi(\lambda)$ values were at the .93 level. The use of .60 cut scores resulted in $\Phi(\lambda)$ figures at the .70 level for one test administration, and .85 for another.

Summary

The primary objective of the present paper was to exemplify the application of *Standards* 2.14, 2.15, and 14.15, deriving and publishing the statistics recommended when cut scores are used for classifying students.

Twelve data sets were employed. Two software systems, Lertap 5 (Nelson, 2000), and BB-Class (Brennan, 2004) were applied to the data sets in order to obtain the statistics wanted: estimates of (1) the standard error of measurement at the cut score (CSEM), (2) classification consistency (p_0), and (3), the value of the dependability index at the cut score ($\Phi(\lambda)$). Results are presented above in Tables 1 and 3.

It is hoped that the publication of these results will contribute to establishing basal data for others to refer to. The twelve data sets represent the work of professional test developers in three areas of the world; they are thought to be representative of present day practice.

A secondary objective was to investigate and experiment with the use of appropriate software. Prior to this paper, Lertap 5 could be used to derive estimates of p_0 and $\Phi(\lambda)$, but it did not have a facility for reporting CSEMs, conditional standard error of measurement estimates. (Now it does.)

The TVAP program (Biddle, 2003) is a comprehensive system; its manual alone represents a solid resource for developers and users of mastery tests. However, its use of Thorndike's (1951) method of deriving CSEMs would not be recommended unless sample size is sufficient to provide an adequate number of results within each score interval.

Lertap computes its estimate of p_0 using a method suggested by Peng and Subkoviak (1980), a computational shortcut to Huyhn (1976). Hanson and Brennan (1990) found reason to question the adequacy of the two-parameter beta binomial model which underpins Huyhn's approach, finding that a four-parameter model was superior in some cases. Brennan's (2004) BB-Class program was used to look into this matter; it was found that the Peng-Subkoviak estimates produced by Lertap were adequate as long as test score distributions were free from clusters of scores at either end of the range.

The appendix provides an example of Lertap 5 control lines which would produce reports with CSEM, p_0, and $\Phi(\lambda)$ estimates for a range of cut-score ("mastery=") settings.

References

AERA/APA/NCME (1999). *Standards for Educational & Psychological Testing.* American Educational Research Association. Washington, D.C.

Berk, R.A. (1980). A consumer's guide to criterion-referenced test reliability. *Journal of Educational Measurement, 17(4)*, 323-349.

Berk, R.A. (2000). *Ask mister assessment person: How do you estimate the reliability of teacher licensure / certification tests?* National Evaluation Systems, Inc: http://www.nesinc.com/PDFs/2000_11Berk.pdf (as at November, 2006).

Biddle Consulting Group, Inc. (2003). *Manual for TVAP: Test validation and analysis Program.* http://www.biddle.com/tvap.stm (as at November, 2006).

Brennan, R.L. (1980). Applications of generalizability theory. In R.A. Berk (Ed.), *Criterion-referenced measurement: The state of the art.* Baltimore, MD: The Johns Hopkins University Press.

Brennan, R.L. (1992). *Elements of generalizability theory (rev. ed.).* Iowa City, IA: American College Testing.

Brennan, R.L. (2001). *Generalizability theory.* New York: Springer.

Brennan, R.L. (2004). *Manual for BB-Class: a computer program that uses the beta-binomial model for classification consistency and accuracy.* University of Iowa: CASMA Research Report Number 9.

Brennan, R.L., & Kane, M.T. (1977). An index of dependability for mastery tests. *Journal of Educational Measurement, 14(3)*, 277-289.

Brown, J.D. & Hudson, T. (2002). *Criterion-referenced language testing.* London: Cambridge University Press.

Crocker, L.M. & Algina, J. (1986). *Introduction to classical and modern test theory.* New York: Holt, Rinehart, and Winston.

Dimitrov, D.M. (2003). Reliability and true-score measures of binary items as a function of their Rasch difficulty parameter. *Journal of Applied Measurement, 4(3),* 222-233.

Feldt, L.S. (1984). Some relationships between the binomial error model and classical test theory. *Educational and Psychological Measurement, 44,* 883-891.

Hanson, B.A., & Brennan, R.L. (1990). An investigation of classification consistency indexes estimated under alternative strong true score models. *Journal of Educational Measurement, 27*, 345-359.

Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement, 13*, 253-264.

Keats, J. A. (1957). Estimation of error variances of test scores. *Psychometrika, 22,* 29-41.

Jarjoura, D. (1986). An estimator of examinee-level measurement error variance that considers test form difficulty adjustments. *Applied Psychological Measurement, 10,* 175-186.

Lord, F.M. (1955). Sampling fluctuations resulting from the sampling of test items. *Psychometrika, 17*, 510-521.

Lord, F.M. (1984). Standard errors of measurement at different ability levels. *Journal of Educational Measurement, 21(3)*, 239-243.

Nelson, L.R. (2000). *Item analysis for tests and surveys using Lertap 5.* Perth, Western Australia: Faculty of Education, Social Work, and Language Studies, Curtin University of Technology. http://www.lertap.curtin.edu.au (at November 2006).

Peng, C-Y.J., & Subkoviak, M.J. (1980). A note on Huynh's normal approximation procedure for estimating criterion-referenced reliability. *Journal of Educational Measurement, 17*, 359-368.

Qualls-Payne, A. (1992). A comparison of score level estimates of the standard error of measurement. *Journal of Educational Measurement, 29,* 213-225.

Shavelson, R., & Webb, N. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage.

Subkoviak, M.J. (1976). Estimating reliability from a single administration of a mastery test. *Journal of Educational Measurement, 13*, 265-276.

Subkoviak, M.J. (1988). A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *Journal of Educational Measurement, 25*, 47-55.

Swaminathan, H., Hambleton, R.K., & Algina, J. (1974). Reliability of Criterion-Referenced Tests: A Decision-Theoretic Formulation. *Journal of Educational Measurement, 11(4),* 263-267.

Thorndike, R.L. (1951). Reliability. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 560-620). Washington, DC: American Council on Education.

Appendix

The Lertap control lines shown below would produce complete test statistics for an 80-item test with cut-score settings ranging from 50% to 90%. In this example, the first item response is found in data column 2; the *key lines would normally have 80 entries. Lertap will produce Statsf, Statsb, CSEM, and Statsul reports for each mastery= setting. The first three of these reports will be identical for each setting, but the Statsul reports will have unique information under their Variance Components section, in the lines titled "Index of dependability" and "Prop. consistent placings".

```
*col(c2-c81)
*sub mastery=50
*key BCAADB……..
*col(c2-c81)
*sub mastery=60
*key BCAADB……..
*col(c2-c81)
*sub mastery=70
*key BCAADB……..
*col(c2-c81)
*sub mastery=80
*key BCAADB……..
*col(c2-c81)
*sub mastery=90
*key BCAADB……..
```