Dave Weiss

~~*S81#2~~

SY81-01

Estimating the Reliability of Adaptive Tests from a Single Test Administration

J. B. Sympson

Educational Testing Service

1981

This manuscript draft has been distributed for review purposes. Please do not cite without obtaining prior permission from the author. Address comments and questions to:

J. B. Sympson
Psychometric Research Group
Educational Testing Service
Princeton, NJ 08541

Abstract

In this paper it is demonstrated that the reliability of a wide variety of adaptive (tailored) tests can be estimated using data from a single test administration. The class of tests considered includes all testing strategies that are based on item response theory and such that the test scores are either maximum likelihood or Bayesian minimum-quadratic-loss estimates of the latent trait θ.

It is shown that for adaptive tests of moderate length (20-35 items) the asymptotic properties of the maximum likelihood estimator and the Birnbaum information measure allow estimation of the reliability of maximum likelihood θ estimates with data from a single empirical test administration. It is also shown that regardless of test length the reliability of Bayesian minimum-quadratic-loss estimates of θ can be determined a priori, without ever actually administering the test.

Acknowledgments

The author appreciates the assistance of John O. Dunn and John Jaskir, both of Educational Testing Service, in conducting the data analyses for this research. Development, simulation, and field administration of the STML adaptive test were completed while the author was Project Director for USAF Research Contract #F33615-77-C-0061 at the University of Minnesota.

Estimating the Reliability of Adaptive Tests from a Single Test Administration

Estimation of the reliability of adaptive tests (a.k.a. tailored tests and inidvidualized tests) using data from a single test administration may not seem possible in view of the fact that different people receive different items in such tests. However, this paper demonstrates that the reliability of a wide variety of adaptive tests can be estimated using data from a single test administration.

The class of tests considered includes all testing strategies that are based on item response theory (a.k.a. item characteristic curve theory and latent trait theory; Lord & Novick, 1968) and such that the test scores are either maximum likelihood or Bayesian minimum-quadratic-loss estimates of the latent trait θ. It is shown that for adpative tests of moderate length (20-35 items) the asymptotic properties of the maximum likelihood estimator and the Birnbaum (1968) information measure allow estimation of the reliability of maximum likelihood θ estimates with data from a single empirical test administration. It is also shown that regardless of test length the reliability of Bayesian minimum-quadratic-loss estimates of θ can be determined a priori, without ever actually administering the test.

The applicability of the procedures described will be demonstrated via an analysis of data obtained in computer simulations of several different testing strategies. In addition, the results obtained for one of the simulated tests will be compared with the results obtained for the

same test when it was actually administered to 495 U.S. Air Force Jet Engine Mechanic Trainees.

It is relevant to note that the procedures presented can also be applied to conventional tests, in which all examinees receive the same test items. The methods suggested provide estimates of the reliability of maximum likelihood and/or Bayesian minimum-quadratic-loss estimates of θ, regardless of the type of testing strategy that gives rise to these estimates. However, with conventional tests estimation of the reliability of maximum likelihood estimates of θ will often require somewhat longer tests than are studied here. This is because moderate length conventional tests typically do not provide unbiased maximum likelihood estimates of θ throughout the entire range of θ in the population of interest. Moderate length adaptive tests, on the other hand, can achieve this goal.

Definitions

We begin by establishing the following definitions:

Strictly parallel tests = tests for which the conditional distributions of test scores, given θ, are identical. (Note that every test is strictly parallel to itself.)

Weakly parallel tests = tests for which the test information function, $I(\theta)$, is identical at all levels of θ. (Samejima, 1977c, p. 194).

Reliability coefficient = Pearson product-moment (PPM) correlation between scores on two strictly parallel tests whose scores are independent at fixed θ (local independence of test scores).

Fidelity coefficient = PPM correlation between test scores and θ (Green, 1976, p, 119).

SSP = strategy-score-population (e.g., SSP reliability coefficient and/or SSP fidelity coefficient).

(In these definitions, and throughout this paper, it is assumed that the tests under consideration measure the same trait.)

We may note that strictly parallel tests are interchangeable from the point-of-view of an examinee. Since the conditional distributions of test scores on such tests are identical at any given θ, the examinee must be indifferent (prior to examining the test content) as to which test she/he is administered. This is not the case for tests whose conditional score distributions have identical lower order (e.g., first and second) moments, since a given examinee may attach great utility or disutility to some particular score that has a different probability of occurrence at her/his ability level on the two tests. Thus, tests that are weakly parallel might be considered interchangeable from the psychometrician's point-of-view, but they are not, in general, interchangeable from the examinee's point-of-view.

Throughout the discussion, I will refer to the SSP reliability coefficient and/or SSP fidelity coefficient of a particular test in order to emphasize the fact that such coefficients, like all test-related correlation coefficients, pertain to a particular item selection strategy and scoring method, when implemented in a particular examinee population (Sympson, 1975). A change in the item selection strategy, the scoring

method, and/or the population of interest will usually change the test reliability and fidelity coefficients.

Theorems

In Appendix A to this paper, the following theorems are proved:

(1) If $X_1 = C + e_1$, $X_2 = C + e_2$, and e_1 and e_2 are independently and identically distributed at fixed C, then $\rho(X_1,X_2) = \eta^2(X_1:C) = \eta^2(X_2:C)$, where $\rho(X_1,X_2)$ is the population PPM correlation between X_1 and X_2 and $\eta^2(X_i:C)$ is the squared correlation ratio for predicting X_i (i = 1,2) from C. (This same result is obtained when e_1 and e_2 are not identically distributed at fixed C, if the first two conditional moments of e_1 and e_2 are identical.)

(2) As the number of test items increases, the SSP reliability coefficient $\rho(\hat{\theta}_1,\hat{\theta}_2) = \eta^2(\hat{\theta}_1:\theta) = \eta^2(\hat{\theta}_2:\theta)$ approaches $\sigma^2(\theta)/\sigma^2(\hat{\theta})$, where $\hat{\theta}_1$ and $\hat{\theta}_2$ are locally independent maximum likelihood estimates of θ obtained from two strictly parallel tests. The marginal variance $\sigma^2(\hat{\theta})$ is the same for both tests, since they are parallel.

(3) As the number of test items increases, $(\sigma^2(\hat{\theta}) - \mu[1/I(\hat{\theta})])$ approaches $\sigma^2(\theta)$, where $I(\hat{\theta})$ is the test information function evaluated at $\hat{\theta}$.

(4) As the number of test items increases, weakly parallel tests that utilize maximum likelihood estimates of θ become strictly parallel.

(5) For a test of any length, $\rho^2(\theta*,\theta) = \sigma^2(\theta*)/\sigma^2(\theta)$, where $\theta*$ is a Bayesian minimum-quadratic-loss estimate of θ.

(6) For any two variables Y and Z, $\rho^2(Z,Y) = \eta^2(Z:Y)\ \rho^2(\mu(Z|Y),Y)$, where $\mu(Z|Y)$ is the conditional mean of Z, given Y, weighted by the marginal density of Y.

Implications of Theorems (1) Through (6)

Theorem 1, above, implies that a test's SSP reliability coefficient is not equal to the test's squared SSP fidelity coefficient unless the regression of test scores on θ is linear. This is a general result that applies to all types of test scores, not just $\hat{\theta}$ and θ^*.

Theorem 2 implies that if the test is long enough, and if $\sigma^2(\theta)$ and $\sigma^2(\hat{\theta})$ can be estimated, the reliability of maximum likelihood estimates of θ can be estimated with the ratio of these two variance estimates.

When one can specify the distribution of θ in the population of interest, $\sigma^2(\theta)$ can be determined analytically or numerically from the θ distribution. In this same situation, $\sigma^2(\hat{\theta})$ can be estimated by executing a computer simulation in which "examinees" are drawn from the specified θ distribution and tested. Thus, when the test is sufficiently long and the population distribution of θ can be specified, the SSP reliability coefficient of maximum likelihood θ estimates can be estimated by executing a single simulated test administration, without actually giving the test.

However, if one can specify the distribution of θ in the population of interest, one should probably be generating Bayesian estimates of θ rather than maximum likelihood estimates. The use of maximum likelihood estimates suggests that the population distribution of θ cannot be

specified. When this is the case, estimation of $\rho(\hat{\theta}_1,\hat{\theta}_2)$ by executing a computer simulation is not possible. Under such circumstances, the result presented as Theorem 3 can be utilized.

After testing an empirical sample of examinees from the population of interest, the quantity $\sim\sigma^2(\theta)$ = (est. $\sigma^2(\hat{\theta})$ - est. $\mu[1/I(\hat{\theta})]$) can be calculated, and then divided by est. $\sigma^2(\hat{\theta})$ in order to obtain the empirical reliability estimate $\sim\rho(\hat{\theta}_1,\hat{\theta}_2)$. An unbiased estimate of $\mu[1/I(\hat{\theta})]$ is provided by the sample mean of the $1/I(\hat{\theta})$ values for the examinees in the sample. (See Birnbaum, 1968, pp. 460-464, and Samejima, 1977b, pp. 234-235, for formulas to use in calculating $I(\hat{\theta})$ under different test models and item-scoring rules.)

The asymptotic relationships described in Theorems 2 and 3 have been utilized previously, but without formal supporting arguments, by Samejima (1977b, p. 243; 1977c, p. 196). These relationships depend, as shown in Appendix A, sections A-2 and A-3, on the fact that $\mu(\hat{\theta}|\theta)$ approaches θ as test length increases. The rapid approach of $\mu(\hat{\theta}|\theta)$ to θ, which is characteristic of the maximum likelihood estimator, allows use of these asymptotic results with adaptive tests of moderate length (20-35 items). See McBride (1975) and Appendix B herein for evidence indicating that θ is essentially unbiased over a fairly wide range of θ in adequately designed adaptive tests of moderate length.

The number of items needed to achieve approximate unbiasedness of $\hat{\theta}$ over the interval of θ of interest will depend on the information properties of the test available and the way items are assigned to examinees during the testing process. In general, tests that use items of high quality and that assign items of appropriate difficulty to each examinee

will not need to be as long as tests that contain poor items and/or make inappropriate item assignments.

The degree to which a given test provides unbiased $\hat{\theta}$ values can be assessed by executing a computer simulation of the testing procedure and observing the mean value of $\hat{\theta}$ at selected θ levels within the interval of θ that will ultimately be tested. This type of simulation does not require one to specify the population distribution of θ. It does require one to identify the interval of θ within which the marginal density $f(\theta)$ is likely to be large enough to have an influence of practical consequence on $\rho(\hat{\theta}_1,\hat{\theta}_2)$. Within this interval, several hundred $\hat{\theta}$ values can be generated at each selected θ level and their mean used as an estimate of $\mu(\hat{\theta}|\theta)$.

For certain non-adaptive tests, the expectation of $\hat{\theta}$ given θ can be derived analytically. In either the computer simulation approach or the analytic approach to estimating/calculating $\mu(\hat{\theta}|\theta)$, some rule for assigning finite values to non-convergent (potentially infinite) maximum likelihood estimates must be specified and the values assigned to such cases included in the analysis.

Theorem 4, above, implies that the asymptotic formula suggested by Samejima (1977c, p. 196) for estimating the PPM correlation between maximum likelihood estimates of θ obtained from any two weakly parallel tests is justified because, in fact, the tests she considers are <u>effectively strictly parallel</u>. This theorem also implies that the SSP reliability coefficient for maximum likelihood estimates obtained from

a test of sufficient length is (approximately) equal to the population correlation of these estimates with maximum likelihood θ estimates obtained from any other test that has an identical test information function.

Thus, another method for estimating the SSP reliability coefficient of maximum likelihood estimates is to actually construct two tests that have $I_1(\theta) \simeq I_2(\theta)$ over the range of θ spanned by the population of interest, administer the two tests to a random sample from the population, and correlate the two sets of θ estimates. This correlation is an estimate of the SSP reliability coefficient for both tests, even if they contain different numbers of items, etc. The obtained value can be compared to the reliability estimates obtained by calculating $[\sim\sigma^2(\theta)/\text{est. } \sigma^2(\hat{\theta})]$ for each test individually.

Theorem 5 implies that the squared SSP fidelity coefficient for Bayesian minimum-quadratic-loss estimates of θ can be estimated with the ratio $[\text{est. } \sigma^2(\theta^*)]/\sigma^2(\theta)$ regardless of test length. $\sigma^2(\theta)$ can be determined analytically or numerically from the Bayesian prior distribution of θ. $\sigma^2(\theta^*)$ can be estimated either by executing a computer simulation in which "examinees" are drawn from the specified prior distribution and tested, or by actually administering the test to a sample of examinees from the population of interest. Again, only one test administration (simulated or live) is required.

Theorem 6, in conjunction with Theorem 1, implies that the squared SSP fidelity coefficient is equal to the unsquared SSP reliability coefficient, multiplied by the squared PPM correlation between the conditional means of the test score and θ. Since both $\hat{\theta}$ and θ^* are asymptotically

unbiased, the squared PPM correlation between their conditional means and θ approaches 1.0, asymptotically. However, the approach to unbiasedness is generally more rapid for the maximum likelihood estimator than for the Bayesian estimator (McBride, 1975; Sympson, 1977). Thus, for any given finite test length, the correlation between conditional means and θ will usually be higher for the maximum likelihood estimator.

The preceding comments notwithstanding, results to be presented next suggest that for tests of moderate length, the correlation of conditional θ^* means with θ is close enough to 1.0 to allow one to use the squared estimated SSP fidelity coefficient as a lower-bound estimate of the SSP reliability coefficient.

Application of Procedures to Simulated Test Data

McBride (1975) executed a computer simulation of a 20-item adaptive test assuming a 3-parameter item characteristic curve (ICC) model. During the test, provisional estimates of θ were generated using Owen's (1975) Bayesian scoring algorithm under the assumption that θ was distributed N(0,1). The ICC difficulty parameter of each item administered was set equal to the provisional Bayesian ability estimate obtained following the preceding item. The ICC discrimination and lower asymptote parameters were set to 1.25 and .20, respectively, for all items. At the end of each 20-item test, the final θ^* value, and a $\hat{\theta}$ value based on the same vector of 20 item responses, were recorded.

McBride simulated 100 adaptive tests at each of 32 θ levels, ranging from -3.20 to +3.00 in .20 steps, and reported the conditional mean and variance of θ^* and $\hat{\theta}$ at each θ level in an appendix table (p. 56). With the exception of the data for $\theta = -3.20$, I used McBride's conditional means and variances to compute an estimate of the squared SSP fidelity coefficient for θ^*, an estimate of the SSP reliability coefficient for θ^*, an asymptotic estimate of the SSP reliability of $\hat{\theta}$, and a non-asymptotic (criterion) estimate of the reliability of $\hat{\theta}$. All of these estimated coefficients were computed under the assumption that θ was distributed N(0,1) in the population of interest.

The results of my analysis of McBride's conditional θ^* statistics are shown in Table 1. The estimate of $E_\theta[\sigma^2(\theta^*|\theta)]$ shown in Table 1

Insert Table 1 about here

---

was obtained by computing a weighted average of 31 of the conditional θ^* variances reported by McBride, and correcting the bias in the resulting value (see Section A-7 of Appendix A, herein). In this computation, each conditional variance was weighted by the area under a standardized normal density function in the interval ranging from .10 below the point at which the conditional variance was generated, to .10 above that point. These normal curve areas were rounded to three decimal places.

The estimate of $V_\theta[\mu(\theta^*|\theta)]$ shown in Table 1 was computed from the (weighted) variance of 31 of the conditional θ^* means reported by McBride, with a further adjustment to correct (approximately) for the bias in the resulting value (see Section A-7 of Appendix A). The estimate of $\mu(\theta^*)$ is a weighted average of the 31 conditional θ^* means, and the estimate of $\sigma^2(\theta^*)$ was set equal to the sum of the first two values in Table 1 (again, see Section A-7 of Appendix A). The estimated value of $\rho^2(\theta^*,\theta)$, the squared SSP fidelity coefficient, is equal to the estimate of $\sigma^2(\theta^*)$, since $\sigma^2(\theta) = 1.0$ in this population.

The estimate of $\eta^2(\theta^*:\theta) = \rho(\theta_1^*,\theta_2^*)$, the SSP reliability coefficient, is equal to the ratio of the estimated value of $V_\theta[\mu(\theta^*|\theta)]$ to the estimated value of $\sigma^2(\theta^*)$. This estimate is the "criterion" estimate of the SSP reliability of θ^* for this test. It is a criterion estimate because, except for the fact that a finite number of observations were made at each θ level and only a finite number of θ levels were observed, it gives the population value of $\rho(\theta_1^*,\theta_2^*)$. Obviously, this estimator cannot be computed in empirical samples. The estimate of $\rho(\theta_1^*,\theta_2^*)$ in Table 1 is seen to exceed the estimate of $\rho^2(\theta^*,\theta)$ by .01, which is consistent with the known relationship between the two unobservable population parameters (recall the previous discussion of the implications of Theorem 6).

Insert Table 2 about here

The results of my analysis of the conditional $\hat{\theta}$ statistics from McBride's simulation are shown in Table 2. The first four parameter estimates in this table were computed as discussed above in connection with Table 1. The asymptotic estimator of $\rho(\hat{\theta}_1,\hat{\theta}_2)$, the SSP reliability coefficient, is equal to the reciprocal of the estimate of $\sigma^2(\hat{\theta})$, since $\sigma^2(\theta) = 1.0$. The non-asymptotic (criterion) estimate, $\eta^2(\hat{\theta}:\theta)$, is equal to the ratio of the second and fourth entries in Table 2 and is seen to be quite close to the asymptotic estimate. While neither of the reliability estimates shown in Table 2 can be computed using data from an empirical sample, these results do serve to suggest that the asymptotic formula derived in Appendix A, Section A-2, can be applied to a well-designed adaptive test that is only 20 items in length.

McBride and Weiss (1976, Study 4, pp. 18-26) conducted another computer simulation of Bayesian adaptive testing in the context of a 3-parameter ICC model. In this simulation, Owen's Bayesian algorithm was used to administer 30 items to each of 100 "examinees" at 31 levels of θ ranging from -3.00 to +3.00 in .20 steps. As before, the estimation procedure assumed that θ was distributed N(0,1). After administering any given item, McBride and Weiss set the ICC difficulty parameter for the next item equal to an "optimal" difficulty level for items with ICC discrimination and lower asymptote parameters of 1.25 and .20, respectively. "Optimal" difficulty in this study was defined as the difficulty level which maximized the item information function (Birnbaum, 1968, pp. 460-464) at the current estimated value of θ.

Once the difficulty of the next item to administer was determined, McBride and Weiss set the item's ICC lower asymptote parameter equal to .20 and the item's discrimination parameter was determined by drawing a random number from a truncated normal distribution with $\mu = 1.25$ and $\sigma = .3$ (before truncation). If the resulting random number was less than .80, the item discrimination parameter was set to .80. Otherwise, the discrimination parameter was set equal to the obtained random number. (This description applies only to the condition simulated by McBride and Weiss in which the item discrimination and difficulty parameters were uncorrelated. Data from McBride and Weiss' "r_{ab}+.71" and "r_{ab}-.71" conditions were not used here.)

McBride and Weiss reported the conditional means and standard deviations of the θ* values generated by their adaptive test in two appendix tables (pp. 33-34). This data was used to obtain the parameter estimates shown in Table 3. The entries in Table 3 were computed in the

---

Insert Table 3 about here

---

same manner as the entries in Table 1. However, contrary to Table 1, the estimated squared SSP fidelity coefficient obtained was slightly larger (approximately .0005) than the criterion estimate of the SSP reliability coefficient, a condition that cannot hold in the population. If more "examinees" had been tested at each θ level, and if a greater number of θ levels had been observed in the McBride and Weiss simulation, this slight reversal of the obtained parameter estimates probably would not

have occurred. We may note that one would expect the two population parameters to be more nearly equal for a 30-item Bayesian test than for a 20-item Bayesian test, since the regression of θ^* on θ is more nearly linear at the longer test length.

Sympson (1979) conducted, along with several other analyses, a computer simulation of a tailored testing strategy he refers to as the "Stratified Maximum Likelihood" (STML) strategy. In the STML strategy, a large item pool is sorted with respect to the value of the item information function at each of several θ levels (e.g., from $\theta = -3.00$ to $\theta = +3.00$ in .25 steps). During the test, the most informative item at the θ level closest to the current value of $\hat{\theta}$ is selected for administration. A good item might appear near the top of several "strata", but once it is administered, it is removed from all strata simultaneously.

Sympson adopted the 3-parameter logistic ICC model (Birnbaum, 1968, p. 405) and simulated 500 administrations of a 35-item STML test at each of 101 levels of θ ranging from -5.00 to +5.00 in .10 steps. The item parameters of the simulated item pool corresponded to parameter values observed in a previous empirical calibration of 280 multiple-choice word knowledge (vocabulary) items. The conditional means and variances of $\hat{\theta}$ that were obtained in the interval from $\theta = -3.00$ to $\theta = +3.00$ are reported in Appendix B herein. The normal curve areas (multiplied by 1,000) that were subsequently used in computing the first three entries in Table 4, and the conditional means of "examinee" $1/I(\hat{\theta})$ values in this interval are also shown in Appendix B.

Insert Table 4 about here

The first six entries in Table 4 were computed in the same manner as the corresponding entries in Table 2, except the $f(\theta)$ values shown in Appendix B were used, since the distance between observed θ values was now .10 rather than .20. The value of the asymptotic estimator of $\rho(\hat{\theta}_1,\hat{\theta}_2)$ is again equal to the reciprocal of the estimate of $\sigma^2(\hat{\theta})$, since $\sigma^2(\theta) = 1.0$, and is seen to exceed the value of the non-asymptotic (criterion) estimate by .007.

The weighted average, over all levels of θ, of the conditional mean of $1/I(\hat{\theta})$ was .047. This quantity was subtracted from the estimated value of $\sigma^2(\hat{\theta})$ in order to obtain the "empirical" estimator of the variance of θ, $\sim\sigma^2(\theta)$. The latter quantity was used to obtain the "empirical" reliability estimate $\sim\rho(\hat{\theta}_1,\hat{\theta}_2)$, which is seen to fall approximately midway between the asymptotic estimator and the non-asymptotic (criterion) estimator of the SSP reliability coefficient.

Examination of Appendix B of this paper provides convincing evidence that in the interval from $\theta = -3.00$ to $\theta = +3.00$ this 35-item test is very nearly unbiased, and at each level of θ the quantity $1/I(\hat{\theta})$ provides a reasonable estimate of $\sigma^2(\hat{\theta}|\theta)$. There is some indication that at this test length $1/I(\hat{\theta})$ has a small negative bias as an estimator of $\sigma^2(\hat{\theta}|\theta)$. In 61 comparisons, the mean of 500 $1/I(\hat{\theta})$ values exceeds $V(\hat{\theta}|\theta)$ only 12 times. Nevertheless, the effect of this

tendency on the "empirical" estimator $\sim\rho^2(\hat{\theta}_1,\hat{\theta}_2)$ is small in the assumed population, as is seen in the closeness of the empirical estimate to the asymptotic estimator and the non-asymptotic (criterion) estimator.

Samejima (1977c, p. 197) reported the results of a computer simulation of a 35 item conventional (non-adaptive) test based on a graded-response ICC model. In this test, all response characteristic functions approached either 0 or 1 as θ decreased toward negative infinity (i.e., there was no "guessing"). The ICC difficulty values of the "best response" categories for the test items were uniformly distributed over the interval from -3.75 to +4.75 in .25 steps. The item discrimination parameters ranged from 1.40 to 2.00 in .10 steps, with 5 items at each level of discrimination.

Samejima simulated the administration of this test to 5 examinees at each of 100 θ levels ranging from -2.475 to +2.475, in .05 steps. Table 5 contains certain parameter estimates reported by Samejima and other parameter estimates derivable from the values Samejima reported. In the analyses leading to Tables 1 through 4, $\sigma^2(\theta)$ was 1.0.

Insert Table 5 about here

In Samejima's simulation, on the other hand, $\sigma^2(\theta)$ was 2.083. Comparison of this value to $\sim\theta^2(\theta)$ in Table 5 suggests that $1/I(\hat{\theta})$ may have tended to underestimate $\sigma^2(\hat{\theta}|\theta)$ somewhat in Samejima's simulation also.

The "empirical" estimator $\sim\rho(\hat{\theta}_1,\hat{\theta}_2)$ differs from the asymptotic estimator by .008 in Samejima's study, which is somewhat larger than the difference of .004 observed in the Sympson (1979) simulation. Unfortunately, Samejima did not report the conditional means and variances of the $\hat{\theta}$ values, so a non-asymptotic (criterion) estimate of $\rho(\hat{\theta}_1,\hat{\theta}_2)$ could not be computed for this simulation.

The estimated SSP reliability coefficient for Samejima's 35-item test is larger than the estimate obtained for Sympson's 35-item STML test, in spite of the close similarity of the two estimated $E_\theta[\mu(1/I(\hat{\theta})|\theta)] = \mu[1/I(\hat{\theta})]$ values (.046 and .047, respectively). This is because different populations are involved in the two simulations. The value of $\sigma^2(\theta)$ was considerably larger in Samejima's population. This demonstrates that whenever SSP reliability coefficients are estimated in two different populations, they are not directly comparable.

While the test that Samejima simulated was non-adaptive, its inclusion in this discussion should not be construed as an argument for the application of the reliability estimation procedure that has been described here to typical 35-item conventional tests that generate maximum likelihood estimates of θ. The test simulated by Samejima was quite unusual for a conventional test in that it had a test information function that was relatively high and virtually constant over the interval from $\theta = -3.00$ to $\theta = +3.00$ (Samejima, 1977a, p. 166). This type of information function is much more typical of well-designed adaptive tests than it is of non-adaptive tests (Vale, 1975; Sympson, 1977, pp. 21-22).

While the procedure described here for estimating the SSP reliability of $\hat{\theta}$ values does not require a constant test information function, it does utilize asymptotic approximations that may not be accurate enough if the level of test information is too low in θ regions where a substantial portion of the population to be tested is located. Application of the procedure described here to typical conventional tests that generate maximum likelihood estimates of θ may require somewhat greater test lengths than have been studied in this paper.

It should also be noted that the parameter estimates presented in Tables 1 through 5 above were primarily influenced by psychometric sampling error (finite test length) and local sampling error (finite number of cases observed at each θ level). The effect of population sampling error (deviation of the sample θ distribution from the population θ distribution) was generally quite small. Thus, the results presented in these tables are only indicative of the "large sample" performance of the reliability estimation procedures described here.

Application to STML Live Testing Data

Sympson (1979) also administered his 35-item STML word knowledge test to 495 U.S. Air Force Jet Engine Mechanic (JEM) Trainees. The (relative) frequency distribution of $\hat{\theta}$ among 489 of these individuals is shown in Figure 1. (Six cases were eliminated because it was suspected

---

Insert Figure 1 about here

---

they were not really trying to do their best.) Table 6 shows summary statistics computed from the original (ungrouped) $\hat{\theta}$ values.

Insert Table 6 about here

The mean and variance of $\hat{\theta}$ shown in Table 6 suggest that the values $\mu(\theta) = .00$ and $\sigma^2(\theta) = 1.0$ that Sympson used in his computer simulation of the STML test do not apply to the population of JEM trainees. (Recall that $\sigma^2(\theta) = 1.0$ implies that $\sigma^2(\hat{\theta}) \geq 1.0$.) However, the skew and kurtosis of the empirical $\hat{\theta}$ distribution are both near zero and examination of the frequency distribution of $\hat{\theta}$ shows the distribution to be unimodal and bell-shaped (i.e., quasi-normal in appearance).

Table 7 contains parameter estimates for empirically estimating the STML test's SSP reliability coefficient in the JEM population. Comparison of the estimated value of $E_\theta[\mu(1/I(\hat{\theta})|\theta)] = \mu[1/I(\hat{\theta})]$ obtained in the live testing session with the corresponding estimate obtained in the

---

Insert Table 7 about here

STML simulation (Table 4) shows the two values to be identical to 3 decimal places (.047). This level of agreement between the two estimates was not anticipated, since the score information function for the STML word knowledge test is <u>not</u> constant over θ levels (Sympson, 1979). As before, the value of $\sim\sigma^2(\theta)$ was obtained by subtracting the estimate of $\mu[1/I(\hat{\theta})]$ from the estimate of $\sigma^2(\hat{\theta})$.

The empirical estimate of the SSP reliability coefficient from the live testing session is seen to be somewhat lower than the value obtained in the STML simulation (.940 vs. 955). However, when the live testing estimate is corrected for restriction of the range of θ (Guilford, 1965, pp. 342-343, Case I), the resulting estimate of the SSP reliability coefficient in a normally distributed population with $\sigma^2(\theta) = 1.0$ is .995, which is identical, to 3 decimal places, to the "empirical" estimate obtained in the computer simulation and quite close to the criterion reliability coefficient obtained in the simulation (.952).

Implications of Data Analyses

The results of the data analyses summarized in Tables 1 through 7 seem to suggest that the following generalizations are warranted:

(1) For adaptive tests that utilize maximum likelihood estimates of θ, if the test is moderately long (20-35 items) and if the available items are of average quality (or better), the sample quantity $[\sim\sigma^2(\theta)/\text{est. } \sigma^2(\hat{\theta})]$ provides a suitable estimate of the SSP reliability coefficient.

(2) It is possible that the quantity $[\sim\sigma^2(\theta)/\text{est. } \sigma^2(\hat{\theta})]$ tends to overestimate the SSP reliability coefficient by a small amount.

(3) If a test that utilizes Bayesian minimum-quadratic-loss estimates of θ is moderately long, the sample estimate of the squared SSP fidelity coefficient provides a reasonable lower-bound estimate of the unsquared SSP reliability coefficient.

(4) If a test that utilizes Bayesian minimum-quadratic-loss estimates of θ is too short, the difference between $\rho^2(\theta^*,\theta)$ and $\rho(\theta_1^*,\theta_2^*)$ in the population may be too large to justify lower-bound estimation of the latter coefficient via estimates of the former coefficient.

In considering the last two generalizations above, it should be kept in mind that, regardless of test length, both $\rho^2(\theta^*,\theta)$ and $\eta^2(\theta^*:\theta) = \rho(\theta_1^*,\theta_2^*)$ can be estimated virtually without error by conducting a computer simulation of just one administration of a Bayesian test. Estimation of $\rho^2(\theta^*,\theta)$ using an empirical sample to obtain an estimate of $\sigma^2(\theta^*)$ will be somewhat less satisfactory, and lower-bound estimation of $\rho(\theta_1^*,\theta_2^*)$ via an empirical estimate of $\rho^2(\theta^*,\theta)$ is least desirable.

Is the Reliability Coefficient a "Dead Concept"?

Samejima (1977c) states that "the reliability coefficient ... is at the mercy of the heterogeneity of the group of examinees, which has nothing to do with the test itself. We can easily make an erroneous test look good by using a heterogeneous group of subjects and obtaining a large value of the 'reliability coefficient.' Similarly, we can make a good test look bad by using a homogeneous group of subjects (p. 196)." These observations lead Samejima (1977b) to state that "reliability is a dead concept in test theory since it differs from one group of subjects to another (p. 243)." Though Samejima did not explicitly mention the SSP fidelity coefficient in these comments, her criticisms must apply

with equal force to this coefficient. Like the SSP reliability coefficient, its magnitude is a function of the variability in θ in the population of interest.

It is clear that under item response theory reliability and fidelity coefficients have been displaced from a central position in test theory. Population-free concepts such as the information function have taken on a role of great importance in the evaluation of testing procedures (Sympson, 1975). But, are reliability and fidelity coefficients really dead concepts ... without merit or useful application?

It seems that the fact that reliability and fidelity coefficients calculated in different populations ("groups" in Samejima's statements quoted above) are not directly comparable, a fact that was noted earlier in this paper, leads Samejima to reject such coefficients altogether. However, this point-of-view does not give adequate consideration to the value of these coefficients as indices for comparing different testing/scoring strategies <u>within</u> a particular population of interest.

Consider a situation in which one of several testing strategies is to be selected for use in a given examinee population. Unless the score information function for one of the tests is higher than the score information functions of all the other tests over the entire range of θ spanned by the population of interest, one cannot declare one test to be superior to the others on the basis of their information functions alone. Since intersecting information functions are more often the rule than the exception (Vale, 1975), we need some other mechanism for

rank-ordering the tests under consideration. (We shall ignore costs and other non-psychometric considerations in this discussion. See Sympson (1975) for a brief survey of a variety of criteria that can be considered in evaluating testing strategies.)

If errors of estimation are considered undesirable throughout the range of θ spanned by the population of interest (i.e., one's goal is measurement of θ rather than classification of individuals into a small number of categories), if one's interest is in optimizing measurement for the great majority of individuals in the population (i.e., one is prepared to sacrifice precision in regions of low θ density, if necessary, in order to improve precision overall), and if a quadratic loss function is reasonable, it is appropriate to select the test that has the largest SSP fidelity coefficient in that population. While the fidelity coefficient measures only the degree of linear association between an estimator and θ, and is insensitive to shifts in origin and scale, this does not constitute a fatal objection to the index.

Since the origin and unit of the θ metric are established through arbitrary constraints imposed during item calibration, any affine transformation of the θ metric is equally acceptable. In general, under the conditions specified, we should use the estimator that correlates most highly with θ, even if the estimator's root-mean-square (RMS) error is large. (However, estimation of response probabilities via the latent trait model will require that ICC parameters be transformed in a manner consistent with the relationship between θ and such an estimator.)

Given the potential usefulness of the SSP fidelity coefficient as an index for selecting among tests to be used within a given population, and given the intimate relationship between the SSP reliability coefficient and the SSP fidelity coefficient that was established via Theorems 1 and 6, we may conclude that it is probably premature to consider reliability to be a dead concept. However, it does make sense to restrict comparisons of estimated reliability and fidelity coefficients to estimates obtained in large random samples from the same population (i.e., samples in which $\sigma^2(\theta)$ is essentially constant). A good way to insure that this condition is satisfied is to draw one large random sample from the population of interest and then assign each of the competing tests to a randomly selected portion of the sample. Better still, if one can obtain repeated measures on each individual $\sigma^2(\theta)$ will be precisely the same for each test under consideration. (However, the possible intrusion of warm-up and/or fatigue effects must be considered in the latter case.)

These comments are not intended to suggest that SSP fidelity and reliability coefficients should be estimated and test/score information functions ignored. Whenever possible, information functions should be computed or estimated. The point to be emphasized is that fidelity and reliability coefficients are far from dead concepts when it comes to evaluating competing testing strategies within a given population. This is especially obvious when it is realized that these coefficients can often be estimated using data obtained in a single test administration, and without knowledge of the test's information function or (in the case of maximum likelihood estimates) the population θ distribution.

Appendix A

A-1: Demonstration that $\rho(X_1,X_2) = \eta^2(X_1:C) = \eta^2(X_2:C)$

It is demonstrated below that the unsquared Pearson product-moment correlation between two variables that consist of a shared common component and individual unique components that are locally independent and identically distributed is equal to the squared correlation ratio for predicting either variable from the shared component.

Define $X_1 = C + e_1$ and $X_2 = C + e_2$, where e_1 and e_2 are independently and identically distributed at given C. No further assumptions regarding the distribution of e_i (i = 1,2) are required. From the foregoing definition, we may conclude that the X_i are also independent and identically distributed at given C.

Thus, $\mu(X_1|C)$, the expectation of X_1 at given C, equals $\mu(X_2|C)$, for all C, and $\sigma^2(X_1|C)$, the variance of X_1 at given C, equals $\sigma^2(X_2|C)$. Note that the marginal expectation of X_1, $\mu(X_1)$, is equal to $E_C[\mu(X_1|C)]$, where the outer expectation is taken over levels of C and each $\mu(X_1|C)$ is weighted by the associated marginal frequency f(C). Similarly, $\mu(X_2) = E_C[\mu(X_2|C)]$. Since C has the same marginal distribution with respect to both X_1 and X_2, and $\mu(X_1|C) = \mu(X_2|C)$ for all C, we conclude that $\mu(X_1) = \mu(X_2)$. Also,

$$\sigma^2(X_1) = E_C[\sigma^2(X_1|C)] + V_C[\mu(X_1|C)]$$

$$= E_C[\sigma^2(X_2|C)] + V_C[\mu(X_2|C)] = \sigma^2(X_2) \quad ,$$

where V_C is the (weighted) variance operator, (Lord and Novick, 1968, p. 263). Thus, we may omit subscripts and write $\mu(X)$ for either variable and $\sigma^2(X)$ for either variable when desired.

From the definition of the Pearson product-moment correlation, we may write

$$\rho(X_1,X_2) = \frac{\mu(X_1X_2) - \mu(X_1)\mu(X_2)}{\sigma(X_1)\sigma(X_2)} \quad . \qquad \text{(A-1.1)}$$

Since $\mu(X_1) = \mu(X_2) = \mu(X)$ and $\sigma(X_1) = \sigma(X_2) = \sigma(X)$, (A-1.1) may be written as

$$\rho(X_1,X_2) = \frac{\mu(X_1X_2) - [\mu(X)]^2}{\sigma^2(X)} \quad .$$

But,

$$\mu(X_1X_2) = E_C[\mu(X_1X_2|C)]$$

$$= E_C[\mu(X_1|C)\ \mu(X_2|C)]$$

$$= E_C([\mu(X|C)]^2) \quad ,$$

since X_1 and X_2 are independent and $\mu(X_1|C) = \mu(X_2|C)$ for any C. Thus,

$$\rho(X_1,X_2) = \frac{E_C([\mu(X|C)]^2) - [\mu(X)]^2}{\sigma^2(X)} \quad .$$

Since $\mu(X) = E_C[\mu(X|C)]$,

$$\rho(X_1,X_2) = \frac{E_C([\mu(X|C)]^2) - (E_C[\mu(X|C)])^2}{\sigma^2(X)}$$

$$= \frac{V_C[\mu(X|C)]}{\sigma^2(X)}$$

$$= \frac{\sigma^2(X) - E_C[\sigma^2(X|C)]}{\sigma^2(X)}$$

$$= 1 - \frac{E_C[\sigma^2(X|C)]}{\sigma^2(X)} \quad , \qquad \text{(A-1.2)}$$

since $\sigma^2(X) = E_C[\sigma^2(X|C)] + V_C[\mu(X|C)]$.

The definition of the correlation ratio (η^2) for predicting any variable, say Z, from another, say Y, is

$$\eta^2(Z:Y) = 1 - \frac{E_Y[\sigma^2(Z|Y)]}{\sigma^2(Z)}$$

(Lord and Novick, 1968, p. 263). Comparison of this definition to (A-1.2) makes it clear that $\rho(X_1,X_2) = \eta^2(X:C) = \eta^2(X_1:C) = \eta^2(X_2:C)$. If, and only if, the regression of X on C is linear, $\eta^2(X:C) = \rho^2(X,C)$ so that $\rho(X_1,X_2) = \rho^2(X,C)$. If the regression of X on C is nonlinear, $\rho(X_1,X_2) > \rho^2(X,C)$.

It is worth noting that these same conclusions are reached if we relax the assumption that e_1 and e_2 are identically distributed at fixed C and assume only that the first two conditional moments of e_1 and e_2 are equal. A review of the derivation presented will show that the key requirements are that e_1 and e_2 are locally independent, $\mu(X_1|C) = \mu(X_2|C)$, and $\sigma^2(X_1|C) = \sigma^2(X_2|C)$.

A-2: Demonstration that $\rho(\hat{\theta}_1,\hat{\theta}_2) \to \sigma^2(\theta)/\sigma^2(\hat{\theta})$

It is demonstrated below that as the number of test items increases, the unsquared SSP reliability coefficient for maximum likelihood estimates of θ approaches $\sigma^2(\theta)/\sigma^2(\hat{\theta})$.

Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be maximum likelihood estimates of θ obtained from two strictly parallel tests and such that $\hat{\theta}_1 = \theta + e_1$ and $\hat{\theta}_2 = \theta + e_2$. We assume that the test data are collected in a way that allows the independence of e_1 and e_2 at given θ. Since the tests that generate $\hat{\theta}_1$ and $\hat{\theta}_2$ are strictly parallel, the distributions of e_1 and e_2 at given θ are identical. Under these conditions, the results of section A-1 apply, and $\rho(\hat{\theta}_1,\hat{\theta}_2) = \eta^2(\hat{\theta}:\theta) = \eta^2(\hat{\theta}_1:\theta) = \eta^2(\hat{\theta}_2:\theta)$.

Thus, we may write

$$\rho(\hat{\theta}_1,\hat{\theta}_2) = 1 - \frac{E_\theta[\sigma^2(\hat{\theta}|\theta)]}{\sigma^2(\hat{\theta})}$$

$$= \frac{\sigma^2(\hat{\theta}) - E_\theta[\sigma^2(\hat{\theta}|\theta)]}{\sigma^2(\hat{\theta})}$$

$$= \frac{V_\theta[\mu(\hat{\theta}|\theta)]}{\sigma^2(\hat{\theta})} \quad .$$

But, since the maximum likelihood estimator is asymptotically unbiased (Birnbaum, 1968, p. 457), as the number of test items increases, $\mu(\hat{\theta}|\theta) \to \theta$ and

$$\rho(\hat{\theta}_1,\hat{\theta}_2) \rightarrow \left(\frac{V_\theta(\theta)}{\sigma^2(\hat{\theta})}\right) = \left(\frac{\sigma^2(\theta)}{\sigma^2(\hat{\theta})}\right) \quad .$$

Note also that $\mu(\hat{\theta}|\theta) \rightarrow \theta$ implies asymptotic linearity of the regression of $\hat{\theta}$ on θ. Thus, $\rho(\hat{\theta}_1,\hat{\theta}_2) \rightarrow \rho^2(\hat{\theta},\theta) = \rho^2(\hat{\theta}_1,\theta) = \rho^2(\hat{\theta}_2,\theta)$, since $\eta^2(\hat{\theta}:\theta) \rightarrow \rho^2(\hat{\theta},\theta)$.

A-3: Demonstration that $(\sigma^2(\hat{\theta}) - \mu[1/I(\hat{\theta})]) \rightarrow \sigma^2(\theta)$

It is demonstrated below that as the number of test items increases, the (positive) difference between the variance of $\hat{\theta}$ and the expectation of $1/I(\hat{\theta})$, where $I(\hat{\theta})$ is the test information function evaluated at $\hat{\theta}$, approaches $\sigma^2(\theta)$.

At any given θ, the maximum likelihood estimator $\hat{\theta}$ is asymptotically normally distributed with mean θ and variance equal to $1/I(\theta)$ (Birnbaum, 1968, p. 457). $I(\theta)$ is the test information function defined for dichotomously scored items by Birnbaum (1968, p. 454) and generalized to other item scoring methods by Samejima (1977b, pp. 234-235).

The asymptotic unbiasedness of the maximum likelihood estimator implies that $\sigma^2(\hat{\theta}) \rightarrow (\sigma^2(\theta) + \sigma^2(\hat{\theta} - \theta))$ as test length increases (Samejima, 1977a, pp. 164-165). Thus, $(\sigma^2(\hat{\theta}) - \sigma^2(\hat{\theta} - \theta)) \rightarrow \sigma^2(\theta)$. Now $\sigma^2(\hat{\theta} - \theta) = E_\theta[\sigma^2(\hat{\theta}|\theta) + (\mu(\hat{\theta}|\theta) - \theta)^2]$ at any test length. As test length increases $\mu(\hat{\theta}|\theta) \rightarrow \theta$ and $\sigma^2(\hat{\theta} - \theta) \rightarrow E_\theta[\sigma^2(\hat{\theta}|\theta)]$. Since $\sigma^2(\hat{\theta}|\theta) = 1/I(\theta)$, asymptotically, $\sigma^2(\hat{\theta} - \theta) \rightarrow E_\theta[1/I(\theta)] = \mu[1/I(\theta)]$. Thus, $(\sigma^2(\hat{\theta}) - \mu[1/I(\theta)]) \rightarrow \sigma^2(\theta)$.

As test length increases, $I(\theta)$ increases without limit. At sufficient finite test length, if the test information function does not change too rapidly in the interval $\theta \pm [9/I(\theta)]^{\frac{1}{2}}$ around a given θ, we have $I(\hat{\theta}) \simeq I(\theta)$, with equality holding asymptotically. Thus, as test length increases, $(\sigma^2(\hat{\theta}) - \mu[1/I(\hat{\theta})]) \rightarrow \sigma^2(\theta)$.

A-4: Demonstration that $I(\theta,\hat{\theta}_1) = I(\theta,\hat{\theta}_2)$ Implies Strict Parallelism

It is demonstrated below that if the test information functions for two tests that utilize maximum likelihood estimates of θ are identical, the tests are strictly parallel, asymptotically.

Birnbaum (1968, p. 453) has defined the information function of a given scoring formula (i.e., the score information function) by the expression

$$I(\theta,X) = [\frac{\partial \mu(X|\theta)}{\partial \mu}]^2 [\sigma^2(X|\theta)]^{-1} \quad .$$

If the test score is a maximum likelihood estimate of θ, then X is replaced by $\hat{\theta}$ in the foregoing equation. Asymptotically, the score information function for $\hat{\theta}$ is equal to $I(\theta)$, the test information function (Birnbaum, 1968, pp. 455-457).

Let us consider two tests that are constructed to have identical test information functions. The tests need not contain the same number of items, need not contain items with the same number of response alternatives, and need not contain items with matched item parameters.

All that is required is that $I_1(\theta) = I_2(\theta)$, where $I_i(\theta)$ is the test information function for test i (i = 1,2). The tests we are considering would be termed weakly parallel tests by Samejima (1977c, p. 194).

Since $I_1(\theta) = I_2(\theta)$ for these tests, $I_1(\theta,\hat{\theta}_1) = I_2(\theta,\hat{\theta}_2)$, asymptotically. This allows us to write the asymptotic equation

$$\left[\frac{\partial \mu(\hat{\theta}_1|\theta)}{\partial \theta}\right]^2 [\sigma^2(\hat{\theta}_1|\theta)]^{-1} = \left[\frac{\partial \mu(\hat{\theta}_2|\theta)}{\partial \theta}\right]^2 [\sigma^2(\hat{\theta}_2|\theta)]^{-1} \quad .$$

Thus,

$$\frac{\left[\frac{\partial \mu(\hat{\theta}_1|\theta)}{\partial \theta}\right]^2}{\left[\frac{\partial \mu(\hat{\theta}_2|\theta)}{\partial \theta}\right]^2} = \frac{\sigma^2(\hat{\theta}_1|\theta)}{\sigma^2(\hat{\theta}_2|\theta)} \quad , \qquad \text{(A-4.1)}$$

asymptotically

Since $\mu(\hat{\theta}_i|\theta) \to \theta$ as test length increases, the partial derivatives on the left side of (A-4.1) both approach 1.0. This implies that $\sigma^2(\hat{\theta}_1|\theta)/\sigma^2(\hat{\theta}_2|\theta) \to 1.0$, which implies that $\sigma^2(\hat{\theta}_1|\theta) \to \sigma^2(\hat{\theta}_2|\theta)$. Thus, at any given θ, the conditional distributions of $\hat{\theta}_1$ and $\hat{\theta}_2$ have identical asymptotic means (i.e., θ), and identical asymptotic variances.

But, as mentioned in section A-3, the asymptotic distribution of $\hat{\theta}$ is a normal distribution function. Since any member of this family of functions is completely determined once its first two moments are specified, we see that the asymptotic conditional distributions of $\hat{\theta}_1$ and $\hat{\theta}_2$ for the tests under consideration must be identical, since their first two moments are identical.

Thus, any two tests that utilize maximum likelihood estimates of θ, and which have $I_1(\theta) = I_2(\theta)$ for all θ, are, asymptotically, strictly parallel tests. For sufficiently long finite length tests, the two tests can be considered to be effectively strictly parallel. Given that the tests are effectively strictly parallel, the results obtained in sections A-1, A-2, and A-3 are applicable.

A-5: Demonstration that $\rho^2(\theta^*,\theta) = \sigma^2(\theta^*)/\sigma^2(\theta)$

It is demonstrated below that the squared SSP fidelity coefficient for Bayesian minimum-quadratic-loss estimates of θ is equal to $\sigma^2(\theta^*)/\sigma^2(\theta)$, regardless of test length.

Let X index an item response vector or a single-valued statistic (i.e., score) derived from the vector. Define $\theta^* = \mu(\theta|X)$. Thus θ^* is a Bayesian minimum-quadratic-loss estimator of θ (Owen, 1975; Sympson, 1977). The squared correlation ratio for predicting θ from θ^* is defined by

$$\eta^2(\theta:\theta^*) = 1 - \frac{E_{\theta^*}[\sigma^2(\theta|\theta^*)]}{\sigma^2(\theta)} \quad . \qquad \text{(A-5.1)}$$

Note that $\eta^2(\theta:\theta^*) = \rho^2(\theta^*,\theta)$, since $\mu(\theta|\theta^*) = \theta^*$ (i.e., the regression of θ on θ^* is linear with unit slope). Thus, substituting ρ^2 for η^2 and rearranging (A-5.1) gives

$$\rho^2(\theta^*,\theta) = \frac{\sigma^2(\theta) - E_{\theta^*}[\sigma^2(\theta|\theta^*)]}{\sigma^2(\theta)}$$

$$= \frac{V_{\theta^*}[\mu(\theta|\theta^*)]}{\sigma^2(\theta)}$$

$$= \frac{V_{\theta^*}(\theta^*)}{\sigma^2(\theta)} = \frac{\sigma^2(\theta^*)}{\sigma^2(\theta)} \quad ,$$

since $\sigma^2(\theta) = E_{\theta^*}[\sigma^2(\theta|\theta^*)] + V_{\theta^*}[\mu(\theta|\theta^*)]$, and $\mu(\theta|\theta^*) = \theta^*$. This result holds regardless of test length.

A-6: Demonstration that $\rho^2(Z,Y) = \eta^2(Z:Y)\ \rho^2(\mu(Z|Y),Y)$

It is demonstrated below that for any two variables Y and Z, the squared Pearson product-moment correlation between Y and Z is equal to the squared correlation ratio for predicting Z from Y, multiplied by the squared Pearson product-moment correlation between the conditional Z means and Y.

We may write $\rho^2(Z,Y)$ in the form

$$\rho^2(Z,Y) = 1 - \frac{E_Y[\sigma^2(Z|Y) + (Z' - \mu(Z|Y))^2]}{\sigma^2(Z)} \quad , \qquad \text{(A-6.1)}$$

where Z' is the least-squares linear regression estimate of Z for given Y. The quantity $(Z' - \mu(Z|Y))$ is the deviation from linear regression at any given Y. Z' is also the least-squares linear regression estimate of $\mu(Z|Y)$ that would be obtained if the $\mu(Z|Y)$ values, each weighted by the associated marginal frequency of Y, were regressed on Y.

The quantity $E_Y[(Z' - \mu(Z|Y))^2]$ is the mean squared residual for linear prediction of $\mu(Z|Y)$ from Y, so we may write

$$\rho^2(\mu(Z|Y),Y) = 1 - \frac{E_Y[(Z' - \mu(Z|Y))^2]}{V_Y[\mu(Z|Y)]} \quad ,$$

which implies

$$E_Y[(Z' - \mu(Z|Y))^2] = [1 - \rho^2(\mu(Z|Y),Y)]\, V_Y[\mu(Z|Y)] \quad . \qquad \text{(A-6.2)}$$

Distributing the expectation operator in (A-6.1) and substitution using (A-6.2) gives

$$\begin{aligned}
\rho^2(Z,Y) &= 1 - \frac{E_Y[\sigma^2(Z|Y)] + [1 - \rho^2(\mu(Z|Y),Y)]\, V_Y[\mu(Z|Y)]}{\sigma^2(Z)} \\
&= \frac{\sigma^2(Z) - E_Y[\sigma^2(Z|Y)] - [1 - \rho^2(\mu(Z|Y),Y]\, V_Y[\mu(Z|Y)]}{\sigma^2(Z)} \\
&= \frac{V_Y[\mu(Z|Y)] - [1 - \rho^2(\mu(Z|Y),Y)]\, V_Y[\mu(Z|Y)]}{\sigma^2(Z)} \\
&= \frac{V_Y[\mu(Z|Y)]\, \rho^2(\mu(Z|Y),Y)}{\sigma^2(Z)} \quad ,
\end{aligned}$$

since $\sigma^2(Z) = E_Y[\sigma^2(Z|Y)] + V_Y[\mu(Z|Y)]$.

But,

$$\eta^2(Z:Y) = 1 - \frac{E_Y[\sigma^2(Z|Y)]}{\sigma^2(Z)}$$

$$= \frac{V_Y[\mu(Z|Y)]}{\sigma^2(Z)} \quad .$$

Thus,

$$\rho^2(Z,Y) = \eta^2(Z:Y)\ \rho^2(\mu(Z|Y),Y) \quad .$$

If, and only if, the regression of Z on Y is linear, $\rho^2(\mu(Z|Y),Y) = 1.0$ and $\rho^2(Z,Y) = \eta^2(Z:Y)$. If the regression of Z on Y is nonlinear, $\rho^2(\mu(Z|Y),Y) < 1.0$ and $\rho^2(Z,Y) < \eta^2(Z:Y)$. In particular, it is possible to construct data configurations in which $\eta^2(Z:Y) = 1.0$, $\rho^2(\mu(Z|Y),Y) = 0.0$, and, thus, $\rho^2(Z,Y) = 0.0$.

A-7: Estimating $E_Y[\sigma^2(Z|Y)]$ and $V_Y[\mu(Z|Y)]$

In this section, an unbiased estimate of $E_Y[\sigma^2(Z|Y)]$, where N values of Z have been observed at each level of the continuous variable Y, is obtained. An approximately unbiased estimate of $V_Y[\mu(Z|Y)]$ is also derived. It is shown that combining these two estimators in order to estimate $\sigma^2(Z)$ is equivalent to estimating $\sigma^2(Z)$ with the sum of the sample values $E_Y[V(Z|Y)]$ and $V_Y[E(Z|Y)]$. For the moment, it is assumed that N values of Z are observed at every possible level of Y. This assumption will be modified later in the development.

At each level of Y, the conditional sample mean, $E(Z|Y)$, and sample variance, $V(Z|Y)$, of Z can be calculated. For any chosen level of Y, $E(Z|Y)$ will vary from sample to sample of size N and will have expectation $\mu(Z|Y)$. The variance of $E(Z|Y)$ will be $\sigma^2(Z|Y)/N$. The sample variance $V(Z|Y)$ will be a biased estimator of $\sigma^2(Z|Y)$, but the bias can be corrected by multiplying $V(Z|Y)$ by $N/(N - 1)$.

Now consider the quantity $E_Y[V(Z|Y)]$. This is just a weighted average of the biased conditional variance estimators, where the subscript on the expectation operator indicates that the expectation is taken over levels of Y. Each conditional variance is weighted by the associated marginal density $f(Y)$. Since the $V(Z|Y)$ are all negatively biased estimators, $E_Y[V(Z|Y)]$ is a negatively biased estimator of $E_Y[\sigma^2(Z|Y)]$. However, correcting the bias in $V(Z|Y)$ at each and every level of Y provides an unbiased estimate of $E_Y[\sigma^2(Z|Y)]$. Since the value of N is constant over Y, the value $N/(N - 1)$ is constant over Y and the correction factor may be taken outside the expectation operation. Thus, $[N/(N - 1)]E_Y[V(Z|Y)]$ is an unbiased estimate of $E_Y[\sigma^2(Z|Y)]$. This shows that $E_Y[V(Z|Y)]$ tends to underestimate $E_Y[\sigma^2(Z|Y)]$.

Now consider the quantity $V_Y[E(Z|Y)]$, where each conditional mean is weighted by an associated $f(Y)$. We may write

$$V_Y[E(Z|Y)] = E_Y([E(Z|Y)]^2) - (E_Y[E(Z|Y)])^2 \quad . \qquad \text{(A-7.1)}$$

Since $E(Z|Y) = \mu(Z|Y) + e$, where e is a local error of estimate,

$$E_Y([E(Z|Y)]^2) = E_Y([\mu(Z|Y) + e]^2)$$

$$= E_Y([\mu(Z|Y)]^2) + 2E_Y[\mu(Z|Y)\ e] + E_Y(e^2) \quad .$$

But,

$$E_Y[\mu(Z|Y)\ e] = E_Y[E([\mu(Z|Y)\ e]|Y)]$$

$$= E_Y[\mu(Z|Y)\ E(e|Y)]$$

$$= E_Y[\mu(Z|Y)\ 0] = 0 \quad ,$$

since $E(Z|Y)$ is an unbiased estimator of $\mu(Z|Y)$. Also,

$$E_Y(e^2) = E_Y[E(e^2|Y)] = E_Y[\sigma^2(Z|Y)/N]$$

$$= (1/N)E_Y[\sigma^2(Z|Y)] \quad ,$$

since

$$\sigma^2(Z|Y)/N = \sigma^2(e|Y) = E(e^2|Y) - [E(e|Y)]^2$$

$$= E(e^2|Y) - [0]^2 = E(e^2|Y) \quad .$$

Thus,

$$E_Y([E(Z|Y)]^2) = E_Y([\mu(Z|Y)]^2) + (1/N)E_Y[\sigma^2(Z|Y)] \quad . \qquad \text{(A-7.2)}$$

The quantity $E_Y[E(Z|Y)] = E(Z)$ is an unbiased estimator of $E_Y[\mu(Z|Y)] = \mu(Z)$, since each $E(Z|Y)$ is an unbiased estimator of its local $\mu(Z|Y)$. Thus, we will utilize the approximation

$$(E_Y[E(Z|Y)])^2 \simeq (E_Y[\mu(Z|Y)])^2 \quad , \qquad \text{(A-7.3)}$$

since E(Z) is unbiased and is based on a very large sample (N times the number of Y levels used) from the bivariate population of (Y,Z) pairs.

With the foregoing approximation at hand, we can substitute (A-7.2) and (A-7.3) into (A-7.1) to obtain

$$V_Y[E(Z|Y)] \simeq E_Y([\mu(Z|Y)]^2) + (1/N)E_Y[\sigma^2(Z|Y)] - (E_Y[\mu(Z|Y)])^2 \quad . \qquad \text{(A-7.4)}$$

Noting that

$$V_Y[\mu(Z|Y)] = E_Y([\mu(Z|Y)]^2) - (E_Y[\mu(Z|Y)])^2 \quad ,$$

and rearranging (A-7.4) gives

$$V_Y[\mu(Z|Y)] \simeq V_Y[E(Z|Y)] - (1/N)E_Y[\sigma^2(Z|Y)] \quad , \qquad \text{(A-7.5)}$$

which indicates that $V_Y[E(Z|Y)]$ tends to overestimate $V_Y[\mu(Z|Y)]$.

$E_Y[\sigma^2(Z|Y)]$ may be replaced by its unbiased estimator $[N/(N-1)]E_Y[V(Z|Y)]$ in (A-7.5) without altering the expectation of the expression over samples. Upon making this substitution, we have

$$V_Y[\mu(Z|Y)] \simeq V_Y[E(Z|Y)] - [1/(N-1)]E_Y[V(Z|Y)] \quad .$$

Thus, $V_Y[E(Z|Y)] - [1/(N-1)]E_Y[V(Z|Y)]$ is an approximately unbiased estimate of $V_Y[\mu(Z|Y)]$. (The completely unbiased estimator of $V_Y[\mu(Z|Y)]$, which is a function of the squared marginal Y densities, is somewhat larger than the estimator suggested here, but still smaller than $V_Y[E(Z|Y)]$.)

In the population,

$$\sigma^2(Z) = E_Y[\sigma^2(Z|Y)] + V_Y[\mu(Z|Y)] \quad .$$

Upon substituting the unbiased estimator of $E_Y[\sigma^2(Z|Y)]$ and the approximately unbiased estimator of $V_Y[\mu(Z|Y)]$ on the right side of this equation, we obtain

$$\text{est. } \sigma^2(Z) = \left(\frac{N}{N-1}\right) E_Y[V(Z|Y)] + V_Y[E(Z|Y)] - \left(\frac{1}{N-1}\right) E_Y[V(Z|Y)]$$

$$= E_Y[V(Z|Y)] + V_Y[E(Z|Y)] \quad .$$

This result indicates that while both $E_Y[V(Z|Y)]$ and $V_Y[E(Z|Y)]$ are biased estimators of their respective parametric values, their biases are (approximately) equal and in the opposite direction.

In the development above, it was assumed that N values of Z are observed at every possible level of Y. If, instead, N values of Z are observed at a number of systematically selected levels of Y, the results obtained above can be applied to the resulting data in order to improve estimation of $E_Y[\sigma^2(Z|Y)]$ and $V_Y[\mu(Z|Y)]$. In order to insure that the estimates will closely approximate the values that would be obtained if one sampled Z values at every level of Y, there should be a large number of closely spaced Y levels actually selected. The levels selected should cover the entire range of Y in which f(Y) is large enough to have an influence of practical consequence on the estimates. At each level of

Y selected, the conditional mean and variance of Z should be weighted in the calculations by the area under f(Y) that is contained in an interval extending from a point halfway between the given level and the next lower level to a point half way between the given level and the next higher level.

Appendix B

Data from STML Word Knowledge Test Simulation

θ	$f(\theta)$	$E(\hat{\theta}\|\theta)$	$V(\hat{\theta}\|\theta)$	$E(1/I(\hat{\theta}))$
-3.00	1	-3.0362	.1435	.1419
-2.90	1	-2.9268	.1320	.1278
-2.80	1	-2.8310	.1308	.1407
-2.70	1	-2.7135	.1101	.1023
-2.60	1	-2.6159	.1009	.0937
-2.50	2	-2.5125	.0860	.0855
-2.40	2	-2.4096	.0870	.0795
-2.30	3	-2.2972	.0751	.0734
-2.20	4	-2.2022	.0788	.0699
-2.10	4	-2.1057	.0663	.0660
-2.00	5	-1.9981	.0689	.0631
-1.90	7	-1.9017	.0613	.0605
-1.80	8	-1.7992	.0691	.0586
-1.70	9	-1.6882	.0601	.0567
-1.60	11	-1.5957	.0574	.0554
-1.50	13	-1.4950	.0594	.0543
-1.40	15	-1.4013	.0590	.0535
-1.30	17	-1.2904	.0563	.0527
-1.20	19	-1.1942	.0529	.0523
-1.10	22	-1.0930	.0564	.0518
-1.00	24	-.9975	.0544	.0514
-.90	27	-.8938	.0523	.0511
-.80	29	-.7866	.0508	.0506
-.70	31	-.6946	.0556	.0501
-.60	33	-.5951	.0496	.0495
-.50	35	-.4964	.0485	.0488
-.40	37	-.3956	.0482	.0478
-.30	38	-.2939	.0491	.0468
-.20	39	-.2027	.0484	.0458
-.10	40	-.1004	.0466	.0448
.00	40	.0011	.0434	.0439
.10	40	.1028	.0477	.0435
.20	39	.2146	.0483	.0434
.30	38	.3127	.0443	.0437
.40	37	.4065	.0469	.0444
.50	35	.5201	.0501	.0451
.60	33	.6135	.0520	.0460
.70	31	.7136	.0510	.0465
.80	29	.8191	.0458	.0472

Appendix B (cont'd.)

θ	$f(\theta)$	$E(\hat{\theta}\|\theta)$	$V(\hat{\theta}\|\theta)$	$E(1/I(\hat{\theta}))$
.90	27	.9144	.0516	.0472
1.00	24	1.0027	.0468	.0471
1.10	22	1.1077	.0487	.0462
1.20	19	1.2059	.0469	.0451
1.30	17	1.2977	.0415	.0436
1.40	15	1.3992	.0414	.0416
1.50	13	1.4971	.0415	.0395
1.60	11	1.5923	.0395	.0373
1.70	9	1.7029	.0361	.0350
1.80	8	1.7955	.0343	.0333
1.90	7	1.8990	.0314	.0317
2.00	5	2.0003	.0298	.0306
2.10	4	2.1042	.0302	.0300
2.20	4	2.2085	.0319	.0298
2.30	3	2.3044	.0290	.0298
2.40	2	2.4099	.0297	.0301
2.50	2	2.5109	.0306	.0303
2.60	1	2.6056	.0326	.0304
2.70	1	2.7074	.0330	.0306
2.80	1	2.8031	.0310	.0307
2.90	1	2.9056	.0309	.0311
3.00	1	3.0073	.0340	.0323

References

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Green, B. F. Invited discussion. In C. L. Clark (Ed.), Proceedings of the First Conference on Computerized Adaptive Testing (U.S. Civil Service Commission, Personnel Research and Development Center, Professional Series 75-6). Washington, D.C.: U.S. Government Printing Office, March 1976.

Guilford, J. P. Fundamental statistics in psychology and education (4th ed.). New York: McGraw-Hill, 1965.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

McBride, J. R. Scoring adaptive tests. In D. J. Weiss (Ed.), Computerized adaptive trait measurement: Problems and prospects (Research Report 75-5). Minneapolis, Minn.: University of Minnesota, Department of Psychology, Psychometric Methods Program, November 1975.

McBride, J. R., & Weiss, D. J. Some properties of a Bayesian adaptive testing strategy (Research Report 76-1). Minneapolis, Minn.: University of Minnesota, Department of Psychology, Psychometric Methods Program, March 1976.

Owen, R. J. A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. Journal of the American Statistical Association, 1975, 70, 351-356.

Samejima, F. A method of estimating item characteristic functions using the maximum likelihood estimate of ability. Psychometrika, 1977, 42 (2), 163-191. (a)

Samejima, F. A use of the information function in tailored testing. Applied Psychological Measurement, 1977, 1 (2), 233-247. (b)

Samejima, F. Weakly parallel tests in latent trait theory with some criticisms of classical test theory. Psychometrika, 1977, 42 (2), 193-198. (c)

Sympson, J. B. Evaluating the results of computerized adaptive testing. In D. J. Weiss (Ed.), Computerized adaptive trait measurement: Problems and prospects (Research Report 75-5). Minneapolis, Minn.: University of Minnesota, Department of Psychology, Psychometric Methods Program, November 1975.

Sympson, J. B. Estimation of latent trait status in adaptive testing procedures. In D. J. Weiss (Ed.), Applications of computerized adaptive testing (Research Report 77-1). Minneapolis, Minn.: University of Minnesota, Department of Psychology, Psychometric Methods Program, March 1977.

Sympson, J. B. Criterion-related validity of conventional and adaptive ability tests in a military environment. Paper presented at the 1979 Computerized Adaptive Testing Conference, Minneapolis, June 1979.

Vale, C. D. Strategies of branching through an item pool. In D. J. Weiss (Ed.), Computerized adaptive trait measurement: Problems and prospects (Research Report 75-5). Minneapolis, Minn.: University of Minnesota, Department of Psychology, Psychometric Methods Program, November 1975.

Table 1

Results of Analysis of McBride Data (Bayesian Scores)

Parameter Estimated	Value Obtained
$E_\theta[\sigma^2(\theta^*\|\theta)]$	.079
$V_\theta[\mu(\theta^*\|\theta)]$	.824
$\mu(\theta^*)$	.010
$\sigma^2(\theta^*)$	.903
$\sigma^2(\theta^*)/\sigma^2(\theta) = \rho^2(\theta^*,\theta)$	.903
$\eta^2(\theta^*:\theta) = \rho(\theta_1^*,\theta_2^*)$	.913

Table 2

Results of Analysis of McBride Data (M.L. Scores)

Parameter Estimated	Value Obtained
$E_{\theta}[\sigma^2(\hat{\theta}\|\theta)]$	.104
$V_{\theta}[\mu(\hat{\theta}\|\theta)]$	1.001
$\mu(\hat{\theta})$	.004
$\sigma^2(\hat{\theta})$	1.105
$\sigma^2(\theta)/\sigma^2(\hat{\theta}) \rightarrow \rho(\hat{\theta}_1, \hat{\theta}_2)$	.905
$\eta^2(\hat{\theta}:\theta) = \rho(\hat{\theta}_1, \hat{\theta}_2)$	.906

Table 3

Results of Analysis of McBride and Weiss Data

Parameter Estimated	Value Obtained
$E_\theta[\sigma^2(\theta^*\|\theta)]$	.072
$V_\theta[\mu(\theta^*\|\theta)]$	.851
$\mu(\theta^*)$	.002
$\sigma^2(\theta^*)$	.923
$\sigma^2(\theta^*)/\sigma^2(\theta) = \rho^2(\theta^*,\theta)$	.923
$\eta^2(\theta^*:\theta) = \rho(\theta_1^*,\theta_2^*)$	.922

Table 4

Results of Analysis of STML Simulation Data

Parameter Estimated	Value Obtained
$E_\theta[\sigma^2(\hat{\theta}\|\theta)]$	.050
$V_\theta[\mu(\hat{\theta}\|\theta)]$	.993
$\mu(\hat{\theta})$	.006
$\sigma^2(\hat{\theta})$	1.043
$\sigma^2(\theta)/\sigma^2(\hat{\theta}) \rightarrow \rho(\hat{\theta}_1,\hat{\theta}_2)$	.959
$\eta^2(\hat{\theta}:\theta) = \rho(\hat{\theta}_1,\hat{\theta}_2)$	.952
$E_\theta[\mu(1/I(\hat{\theta})\|\theta)]$	.047
$\sim\sigma^2(\theta)$	.996
$\sim\sigma^2(\theta)/\sigma^2(\hat{\theta}) = \sim\rho(\hat{\theta}_1,\hat{\theta}_2)$	.955

Table 5

Results Reported by Samejima (1977c)

Parameter Estimated	Value Obtained
$\sigma^2(\hat{\theta})$	2.148
$\sigma^2(\theta)/\sigma^2(\hat{\theta}) \to \rho(\hat{\theta}_1, \hat{\theta}_2)$	.970
$E_\theta[\mu(1/I(\hat{\theta})\|\theta)]$	.046
$\sim\sigma^2(\theta)$	2.102
$\sim\sigma^2(\theta)/\sigma^2(\hat{\theta}) = \sim\rho(\hat{\theta}_1, \hat{\theta}_2)$	.978

Table 6

Summary Statistics for STML Word Knowledge Ability Estimate Distribution

$\hat{\theta}$ Statistic	Value
Mean	-.170
Variance	.788
Skew	-.010
Kurtosis	-.043
Minimum	-2.559
Maximum	2.620
No. of Cases	489

Table 7

Results of Analysis of STML Live Testing Data

Parameter Estimated	Value Obtained
$\mu(\hat{\theta})$	-.170
$\sigma^2(\hat{\theta})$	.788
$E_\theta[\mu(1/I(\hat{\theta})\|\theta)]$	.047
$\sim\sigma^2(\theta)$	.741
$\sim\sigma^2(\theta)/\sigma^2(\hat{\theta}) = \sim\rho(\hat{\theta}_1,\hat{\theta}_2)$	.940
$\sim\rho(\hat{\theta}_1,\hat{\theta}_2)$ if $\sigma^2(\theta) = 1.0$	.955