Estimating Reliabilities of Computerized Adaptive Tests

D. R. Divgi
Center for Naval Analyses

This paper presents two methods for estimating the reliability of a computerized adaptive test (CAT) without using item response theory. The required data consist of CAT and paper-pencil (PP) scores from identical or equivalent samples, and scores for all examinees on one or more covariates. Multiple R^2s and communalities are used to compute the ratio of CAT and PP reliabilities. When combined with the PP reliability calculated by a conventional procedure, these ratios yield estimates of CAT reliability. *Index terms: computerized adaptive testing, item response theory, predictive validity, reliability, tailored testing.*

Item response theory (IRT) and the availability of powerful portable computers have led to the development of computerized adaptive testing (CAT) versions of some test batteries. In CAT, each administered item is selected by a computer using the best available estimate of the examinee's ability or trait level. As a result, different examinees receive different sequences of items; hence classical reliability estimators based on item responses cannot be used. (For more information about CAT and its uses see Weiss, 1982; Weiss & Kingsbury, 1984.)

Estimating Predictive Validity of CAT

Consider a practical problem. For decades, a test battery has been used in the conventional paper-pencil (PP) mode. Its purpose is to predict success in some activity, such as college education or a job. Before PP testing is replaced by CAT, it is necessary to be reasonably confident that the predictive validity of CAT is at least equal to that of the PP version. How can this be done?

The direct way is to administer CAT to a large number of examinees, use it for selection of applicants, and when measures of success are available, correlate them with CAT scores. This direct procedure is inconvenient because it requires a large sample and a few years for the criterion scores to be available.

A simpler procedure is based on classical test theory. Let Y_C and Y_P be scores on CAT and PP versions of a given test, T_C and T_P the corresponding true scores, and X the criterion score. The correlations of true and observed scores with the criterion are related by

$$r(T_P,X) = \frac{r(Y_P,X)}{(\rho_P)^{1/2}} \quad , \tag{1}$$

where ρ_P is the reliability of the PP test (Lord & Novick, 1968, p. 70). As Lord and Novick's proof showed, the only required assumption is that error scores have zero correlations with true scores and with other error scores. Error variance need not be homogeneous (i.e., independent of true score), and the relationship between test and criterion scores need not be linear. For CAT scores

$$r(T_C,X) = \frac{r(Y_C,X)}{(\rho_C)^{1/2}} \quad . \tag{2}$$

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 13, No. 2, June 1989, pp. 145–149
© Copyright 1989 Applied Psychological Measurement Inc.
0146-6216/89/020145-05$1.50

Assuming that the CAT and PP true scores are linearly related, the correlations in Equations 1 and 2 are identical. Thus

$$r(Y_C,X) = r(Y_P,X)\left(\frac{\rho_C}{\rho_P}\right)^{1/2} \quad . \qquad (3)$$

Therefore, the predictive validity of CAT can be computed from that of the PP test and the ratio of their reliabilities. The assumption of linearly related true scores should be well approximated in reality if CAT items measure the same trait as PP items, and CAT scores have been converted to the PP metric by some equating procedure. Any major norm-referenced testing program will ensure both of these conditions; otherwise it will be difficult to justify using the CAT version as a replacement for the PP test.

It is possible that the traits measured by the CAT and PP versions are not quite the same; this possibility can be investigated using structural analysis (Cudeck, 1985). Such an analysis may show that the traits differ, perhaps due to a method factor. This invalidates Equation 3, because the correlation of the criterion with the method factor is unknown and thus CAT validity requires collecting criterion data on CAT examinees. A quick estimate can only be obtained by estimating CAT reliability and using Equation 3. This paper shows how CAT reliability can be estimated without IRT, assuming only that Equation 3 holds for all measures X.

Who Needs Reliability?

With the advent of IRT, which provides item parameters and information functions that are independent of the population of examinees, the appropriateness of the conventional reliability coefficient has been questioned. According to Samejima (1977, p. 243), reliability is a ''dead concept'' because its value depends on the heterogeneity of the group. According to Green, Bock, Humphreys, Linn, and Reckase (1984), ''such an index is somewhat contrived. Many psychometricians feel that devising a reliability coefficient for an adaptive test is inappropriate and misguided'' (p. 352).

No doubt, a plot of conditional error variances at different abilities provides much more information than a single number. However, global indices such as reliability are useful, and sometimes indispensable, depending on the goals and needs of those developing the CAT version of a test. The reliability coefficient is important because it provides the only simple way to estimate the predictive validity of CAT. Although reliability of a test depends on the population, so does its validity.

According to Green et al. (1984), ideally the scores on CAT and PP versions of the same test ''should correlate as well as their reliabilities will allow. Any important reduction from this ideal indicates some change in the nature of the test, and hence the need for revalidation'' (p. 353). Thus, computing the reliability coefficient is neither inappropriate nor misguided, but instead an important part of evaluating the CAT version.

It follows from Equation 3.9.3 in Lord and Novick (1968) that the highest correlation allowed by the reliabilities is $(\rho_C\rho_P)^{1/2}$, provided that the true scores on the two versions are perfectly correlated.

Green et al. recommended computing reliability using the conditional error variance (at fixed θ) obtained from IRT. Because different examinees receive different items in CAT, the conditional error variance cannot be calculated analytically; computer simulation is required. If only reliability is needed without the entire information function, real data can be used (Sympson, 1980). Both methods require knowing the distribution of θ in the population. More important, both require faith in all the assumptions of IRT. In particular, it is necessary to assume that item parameters are known. When the CAT version of a test is introduced, item parameters are estimated from a PP administration. Strong evidence exists (Ackerman, 1985; Divgi, 1986) that parameters of items can change, often substantially, from PP to CAT administration. Therefore, a reliability coefficient computed using IRT is likely to be too large.

Method

This paper presents two methods of estimating CAT reliability without using IRT. The only assumption required is that Equation 3 holds for every variable X. It is necessary for each examinee to have scores on one or more covariates. The CAT and PP versions may be administered to the same group of examinees or to equivalent samples from the same population. The methods are illustrated using data on the Armed Services Vocational Aptitude Battery (ASVAB; Maier & Sims, 1986).

When CAT is implemented for operational use with the ASVAB, its scores will have been transformed to the PP metric. Therefore, as a preliminary step, CAT was equated to PP using the equipercentile procedure. The equated scores, rounded to one decimal, were used in all later analyses. Equating puts CAT scores on the same metric as PP scores, so the assumption of perfectly correlated true scores is reasonable.

The primary assumption is that CAT and PP versions of a subtest measure the same trait. It is supported by Cudeck's (1985) analysis of four ASVAB subtests. He concluded that "the two testing methods are virtually identical" (p. 319). If Equation 3 holds for each covariate, it is also valid when the bivariate correlations are replaced by multiple correlations of CAT and PP scores with the entire set of covariates, giving

$$\frac{\rho_C}{\rho_P} = \frac{R_C^2}{R_P^2} \quad . \qquad (4)$$

Thus, the ratio of CAT and PP reliabilities for any test can be found by regressing both scores on the available covariates. It is useful, but not theoretically necessary, that the covariates have strong correlations with the test.

A ratio of reliabilities can also be obtained using factor analysis. The communality of a score is the proportion of its variance explained by the underlying factors. Thus, it is the squared multiple correlation when the score is regressed on the common factors rather than on observed covariates. Hence, using the same arguments as above, the ratio of reliabilities is given by

$$\frac{\rho_C}{\rho_P} = \frac{h_C^2}{h_P^2} \quad , \qquad (5)$$

where h^2 represents communality.

CAT reliabilities are obtained by multiplying Equation 4 or Equation 5 by the reliability of the PP version. If item data for the PP version are available, a split-half reliability or KR-20 can be calculated. Otherwise some approximation must be used.

Illustration

The ASVAB is used by the military services for selection and classification of enlisted personnel. It contains 10 subtests: General Science (GS), Arithmetic Reasoning (AR), Word Knowledge (WK), Paragraph Comprehension (PC), Numerical Operations (NO), Coding Speed (CS), Auto and Shop Information (AS), Mathematics Knowledge (MK), Mechanical Comprehension (MC), and Electronics Information (EI). NO and CS, being speeded tests, cannot be adaptive. AS has been split in the CAT version into separate Auto and Shop subtests. Hence these three subtests were excluded from the analysis.

The data came from a study of the predictive validity of an experimental CAT version of the ASVAB (Vicino & Hardwicke, 1984). Each recruit was administered CAT-ASVAB plus three to five subtests of PP-ASVAB. In addition, the pre-enlistment scores on all ASVAB subtests were taken from the recruit's files. These served as covariates. For any given subtest, the usable sample consisted of recruits who were administered the PP version of the subtest. The size of this sample varied from one subtest to another. CAT item responses were scored using Owen's (1975) approximation for the Bayesian ability estimate. The order of administration, CAT before PP or vice-versa, was ignored because its effect has been found to be small (Stoloff, 1987).

Communalities were estimated iteratively using the Statistical Analysis System (SAS Institute, 1986). Following earlier factor analyses of the ASVAB (Stoloff, 1983), three factors were extracted. (The speed factor was absent because speeded subtests were excluded.)

Because item data on the PP subtests were not available, their reliabilities were computed using Lord's (1965) compound binomial model. In this model the parameter k, which represents the spread of difficulties among items in the subtest, is treated as a property of the subtest itself, irrespective of the sample to which it is administered. For each subtest, *k* was estimated from the mean, variance, and KR-20 reliability in a national sample tested in 1980 (Maier & Sims, 1986, Tables 2-2 and 4-1). It was then used to estimate reliability in the recruit sample.

Results

Table 1 presents results of factor and regression analyses. For any given subtest, CAT and PP scores have the same mean and standard deviation as a result of equipercentile equating. With the exception of MK and MC, CAT surpasses PP in both communality and multiple correlation, which means the CAT version is more reliable.

Table 2 presents reliability estimates. For the PP version, the estimate based on Lord's model uses only the mean and variance from the recruit sample; it is independent of the results of factor and regression analyses. When combined with the ratios in Equations 4 and 5, it yields the estimates for CAT.

Discussion

While it would have been preferable to compute KR-20 directly from item responses, use of the compound binomial model is not a serious limitation. Even if Lord's parameter *k* is not strictly invariant across populations, the effect is bound to be small. The value of KR-21, which depends only on mean and variance, is a lower bound for KR-20. What depends on the value of *k* is the difference between KR-21 and estimated KR-20, which is small. For example, for PC, KR-21 = .603 and KR-20 = .633. Therefore, the reliability estimate is robust against any changes in *k* from the 1980 sample to the recruit sample.

Ideally, whether for the PP or CAT version, reliability should be determined as the correlation between equivalent forms. However, because this imposes a burden on the examinees, internal consistency indices such as KR-20 are generally used for PP tests. The methods in this paper are presented in the same spirit, for use when it is not feasible to test examinees twice using two CAT forms. Although the recruits in the CAT-ASVAB validity study took both CAT and PP versions, this is not strictly necessary. The two versions can be administered to different but equivalent samples. The crucial ingredient is the availability of covariates. Multiple regression seems preferable to factor analysis because it is conceptually and computationally simpler. Factor analysis requires the psychometrician to make some subjective choices—in particular, the factoring method and the number of factors.

The assumptions involved in these calculations are neither new nor particularly strong. The assumption of uncorrelated errors is standard in test theory, and is much weaker than the IRT assumption of conditional independence among items. The assumption of perfectly correlated true scores has been used by Lord (1973), and was implied by Green et al.'s (1984) criterion for CAT and PP versions measuring the same trait. Of course, some departure from the assumption always occurs in

Table 1
Sample Statistics for Paper-Pencil (PP)
and CAT Subtests

Sub-test	Sample Size	Mean	SD	R^2 PP	R^2 CAT	Communality PP	Communality CAT
GS	1083	18.7	3.64	.613	.717	.649	.751
AR	1613	19.7	5.75	.678	.719	.731	.786
WK	2635	27.5	4.95	.671	.707	.756	.801
PC	2635	11.0	2.60	.367	.484	.401	.541
MK	814	16.2	5.19	.697	.691	.805	.760
MC	2103	15.7	4.57	.579	.570	.647	.614
EI	1021	13.9	3.53	.644	.706	.683	.726

Table 2
Reliability Estimates for Paper-Pencil (PP) and CAT Subtests

		CAT	
Subtest	PP	Regression	Factor
GS	.727	.850	.841
AR	.848	.899	.911
WK	.809	.852	.857
PC	.633	.835	.853
MK	.842	.835	.795
MC	.772	.760	.733
EI	.730	.802	.776

real data; the question is how large the departure is. But if Equation 3 is rejected, not only are the reliability estimators in this paper invalidated, but the only simple way of estimating the predictive validity of CAT is also questioned. In fact, if the correlation between CAT and PP true scores is substantially below unity, that means the CAT version measures an appreciably different trait than the PP version. As a result, the former may be unacceptable as a replacement for the latter.

References

Ackerman, T. A. (1985, October). *An investigation of the effect of administering test items via the computer.* Paper presented at a meeting of the Midwest Educational Research Association.

Cudeck, R. (1985). A structural comparison of conventional and adaptive versions of the ASVAB. *Multivariate Behavioral Research, 20,* 305–322.

Divgi, D. R. (1986). *Determining the sensitivity of CAT-ASVAB scores to changes in item response curves with the medium of administration* (Research Memorandum 86-189). Alexandria VA: Center for Naval Analyses.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21,* 347–360.

Lord, F. M. (1965). A strong true-score theory, with applications. *Psychometrika, 30,* 239–270.

Lord, F. M. (1973). Testing if two measuring procedures measure the same dimension. *Psychological Bulletin, 79,* 71–72.

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

Maier, M. H., & Sims, W. H. (1986). *The ASVAB score scales: 1980 and World War II* (Report 116). Alexandria VA: Center for Naval Analyses.

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

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

SAS Institute, Inc. (1986). *Statistical Analysis System* (Version 5.03). Cary NC: Author.

Stoloff, P. H. (1983). *A factor analysis of ASVAB form 8a in the 1980 DoD reference population* (Memorandum 83-3135). Alexandria VA: Center for Naval Analyses.

Stoloff, P. H. (1987). *Equivalent-groups versus single-group equating designs for the Accelerated CAT-ASVAB Project* (Research Memorandum 87-6). Alexandria VA: Center for Naval Analyses.

Sympson, J. B. (1980, April). *Estimating the reliability of adaptive tests from a single test administration.* Paper presented at the annual meeting of the American Educational Research Association, Boston.

Vicino, F. L., & Hardwicke, S. B. (1984, April). *An evaluation of the utility of large scale computerized testing.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6,* 473–492.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21,* 361–375.

Acknowledgments

The author thanks Bruce Bloxom, William H. Sims, and J. Bradford Sympson for helpful comments, and the Navy Personnel Research and Development Center for providing the CAT-ASVAB dataset.

Author's Address

Send requests for reprints or further information to D. R. Divgi, Center for Naval Analyses, 4401 Ford Avenue, Alexandria VA 22302-0268, U.S.A.