ED 304 477	TM 012 876
AUTHOR	Bekhuis, Tanja C. H. M.
TITLE	The Estimation of True Scores for Tests Not Taken: A Simulation Study.
PUB DATE	Feb 88
NOTE	20p.; Paper presented at the Annual Meeting of the Eastern Educational Research Association (Miami Beach, FL, February 24-27, 1988).
PUB TYPE	Reports - Research/Technical (143) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Achievement Tests; *Computer Simulation; *Estimation (Mathematics); *Latent Trait Theory; *Scoring Formulas; Secondary Education; Simulation; *Standardized Tests; *True Scores
IDENTIFIERS	Educational Testing Service; *LOGIST Estimation Procedures; Number Right Scoring

ABSTRACT

An Educational Testing Service (ETS) procedure was evaluated, which is based on item response theory and estimates true scores on tests not taken. The reading, vocabulary, and mathematics tests of high school seniors from the National Longitudinal Study (NLS) of 1972 and the High School and Beyond (HSB) seniors of 1980 and 1982 were found to share common blocks of items but differ in terms of total test length and overall difficulty. As the common blocks were too short to permit reliable comparisons, LOGIST was used to simultaneously estimate the item and person parameters for all three cohorts. The estimate of a student's ability (theta) based on the achievement test taken was combined with the item parameters for a test not taken. Thus, a number-right true score or "estimated number-right score" was obtained for a 1982 HSB senior on a 1972 NLS mathematics test. Since the group's expected true score equaled its expected number-right score, the desired cohort comparisons across time could be made. In an evaluation of this method, which generates hypothetical true scores based on tests not actually taken, comparable test forms (X and Y) of varying common block size were simulated for three conditions. Two hundred simulated examinees each for the three conditions were used. For each simulated examinee, the probability of a correct response was computed for each item. Comparisons of simulated true scores with true scores based on LOGIST estimates were made for both the tests taken and the hypothetical tests not taken. Results indicate that the ETS method should remain experimental. (TJH)

The Estimation of True Scores for Tests Not Taken: A Simulation Study

Tanja C.H.M. Bekhuis

L.L. Thurstone Psychometric Laboratory

Department of Psychology

University of North Carolina at Chapel Hill

Paper presented at the annual meeting of the

Eastern Educational Research Association

Miami Beach, Florida

February 27, 1988

The Estimation of True Scores for Tests Not Taken

The purpose of this paper is to evaluate an Educational Testing Service (ETS) procedure described by Pollack at the 1985 annual AERA meeting. The ETS procedure is based on item response theory (IRT) and allows the estimation of true scores on tests not taken. In an effort to investigate the apparent decline in achievement test scores, ETS researchers compared high school seniors from the National Longitudinal Study (NLS) of 1972 with the High School and Beyond (HSB) seniors of 1980 and 1982. The reading, vocabulary, and mathematics tests shared common blocks of items but differed with respect to total test length and overall difficulty. Since the common blocks were too short to permit reliable comparisons, ETS used LOGIST (1982) to simultaneously estimate the item and person parameters for all three cohorts. Thus, the estimate of a student's ability (theta) based on the achievement test the student had taken could be combined with the item parameters for a test the student had not taken. In this way, a number-right true score or "estimated number-right score" could be obtained for, say, a 1982 HSB senior on a 1972 NLS mathematics test (Pollack, 1985, p.10). Since the group's expected true score is equivalent to its expected number-right score (Lord & Novick, 1968), the desired cohort comparisons across time could be made. Clearly, a method which generates hypothetical true scores based on tests not actually taken needs to be evaluated.

Method

The author designed a simulation study where LOGIST-based true score estimates could be compared to simulated true scores with respect to overall bias, relative variability, and linear association. Comparisons were made for several alternate forms of a test, as well as an equivalent form for each of three conditions. The model and experimental design will be presented below.

Model:

The three-parameter logistic model specifies the probability that a person of a given ability (θ_k) will correctly answer a dichotomously scored item:

$$P_i(\theta_k) = c_i + (1-c_i)/(1+e^{-1.7a_i(\theta_k-b_i)}) \qquad (1)$$

where $P_i(\theta_k)$ = the probability of a correct response to the i^{th} item by the k^{th} person

c = pseudo-chance score level

e = 2.71828 ...

a = discriminating power of the item, and

b = the item difficulty, a location parameter relative to the theta scale (Birnbaum in Lord, 1980, p.12) .

Some of the major testing companies are utilizing Rasch-like three-parameter logistic models where the discrimination parameter 'a' is fixed, the pseudo-guessing parameter 'c' is fixed, and the item difficulty parameter 'b' is free to vary. This model is "Rasch-like" since only the b_i's are varying. However, once the pseudo-guessing value is specified, both one-parameter or Rasch models, and two-parameter models are strictly ruled out.

Simulated Tests:

Comparable test forms (X and Y) of varying common block size were simulated for three conditions. In each case, $n_x=n_y=20$ items where $a=1.0$ and $c=.2$ for all items. The following item parameters were input to the simulation phase and not estimated:

<u>Condition</u> <u>I</u> (75% of the items are shared by tests X and Y) :

$-1.5 <= b_x <= 1.0$	$\bar{b}_x = -.270$	$SD_{b_x} = .615$
$-1.3 <= b_y <= 1.0$	$\bar{b}_y = -.120$	$SD_{b_y} = .616$

<u>Condition</u> <u>II</u> (50% of the items shared) :

$-1.5 <= b_x <= 1.0$	$\bar{b}_x = -.270$	$SD_{b_x} = .615$
$-1.3 <= b_y <= 1.0$	$\bar{b}_y = -.148$	$SD_{b_y} = .623$

<u>Condition</u> <u>III</u> (No items shared):

$-1.5 <= b_x <= 1.0$	$\bar{b}_x = -.270$	$SD_{b_x} = .615$
$-1.3 <= b_y <= 1.0$	$\bar{b}_y = -.155$	$SD_{b_y} = .684$

Test X is the same across conditions, while test Y differs slightly; test X is always somewhat easier than test Y.

The item parameters are modifications of an array used by Yen (1984, p.96) in a simulation study.

<u>Simulated</u> <u>Examinees</u>:

Two hundred simulated examinees for each condition were obtained by using the RANNOR function in SAS (1982) which randomly generates an observation from a normally distributed population with mean equal to zero and variance equal to one. Each set of thetas (examinees) was rescaled to (0,1) to adjust for sampling fluctuations. This rescaling is necessary for some comparisons of simulated thetas vs thetas based on LOGIST estimates since the theta scale is indeterminate and needs to be fixed (Lord, 1980, p.36). By default, LOGIST standardizes theta to (Ø,1) in two of the four estimation steps and adjusts the item parameters accordingly (LOGIST, 1982, p.15).

Simulated Response Vectors:

For each simulated examinee, the $P_i(\theta_k)$ was computed for each item. To generate a response vector for that person over all items, the SAS (1982) RANUNI function was called. This function returns a uniform deviate on the interval (Ø,1). A response was coded as 1 = correct if the random number from RANUNI was less than or equal to $P_i(\theta_k)$, and Ø = incorrect otherwise. The resultant matrix of responses (U) was 2ØØx4Ø . The first 1ØØ examinees were considered to be Group I, the second 1ØØ examinees as Group II. Thus, the matrix of responses could be partitioned into quadrants (see Figure 1) . Quadrants (ii) and (iii) correspond to the hypothetical tests or the tests not taken and are later coded as 'not reached' for LOGIST. The input dataset for LOGIST has missing data in these quadrants and parallels the case where Group I took test X and Group II took test Y. However, the data corresponding to the tests not taken are saved from the simulation phase for eventual comparisons of simulated vs estimated true scores.

Insert Figure 1 about here.

A matrix of responses was generated for each of the three conditions where the relative sizes of the common blocks of items were 75%, 5Ø%, and no common block for conditions I, II, and III, respectively.

True Scores for Simulated Tests:

The rows of U within a test were summed to obtain number-right scores. The p-values ($P_i(\theta_k)$) for each examinee over the items of a given test were summed to obtain number-right true scores since

$$\xi_k = \sum_{i=1}^{n} P_i(\theta_k) \qquad (2)$$

where ξ_k = the k^{th} person's number right true score (Lord, 198Ø, pp. 45-46).

LOGIST Estimation:

The input data matrix for LOGIST was a transformed U matrix, where U has been described above. Each simulated examinee's response vector consisted of 1's, Ø's, and 3's where 1=correct, Ø=incorrect, and 3=not reached. Thus, an estimated theta is based on the responses to items on the test taken by the examinee -- an examinee is not penalized for the items coded as 'not reached.' Although the estimation of item parameters for both tests is simultaneous and has been referred to by others as "concurrent calibration" (cf. Petersen, Cook, & Stocking, 1983), items for test X are, in effect, calibrated on Group I since noone from Group II took test X, at least as coded for LOGIST. Similarly, items for test Y are calibrated on Group II. The LOGIST input dataset (for each condition) is depicted in Figure 2.

Insert Figure 2 about here.

The limits of the theta scale were constrained to be ± 3.Ø since most practitioners would ordinarily use these limits. Since b's and theta's needed to be estimated (a's and c's were fixed), only the first estimation step was invoked. The convergence criterion was set to the default criterion value normally associated with the fourth and final step of a full LOGIST run (LOGIST, 1982, pp.13-14).

The advantage of inputting the data as described above, is that the simultaneous estimation of person parameters and item parameters over the two groups results in estimates on a common scale (Pollack, 1985, p.9; Hambleton & Swaminathan, 1985, p.212).

Since LOGIST cannot return finite maximum likelihood estimates of theta for examinees with perfect scores, it was necessary to look at the distributions of estimated thetas from each condition to select an appropriate theta value for these examinees. Simulating the case where a practitioner would be forced to assign a theta value, one tenth of a unit was added to the rounded, largest estimated theta so that the distribution would remain relatively continuous. These values were then used in subsequent analyses.

The person and item parameter estimates were later used to obtain p-values and true scores. In order to obtain true scores on the tests not taken, the theta estimates from the group of interest were combined with the item parameters on the test not taken by that group to generate p-values which were then summed. Similarly, to obtain true scores on tests taken the theta estimates from the group of interest were combined with the item parameters on the test taken by that group.

Schematically, the blocks of true scores may be envisioned as in Figure 3.

Insert Figure 3 about here.

True Score and Theta Comparisons:

Comparisons of simulated true scores with true scores based on LOGIST estimates were made for both the tests taken and the hypothetical tests not taken. Several statistics were computed: (1) the standardized mean difference (where the denominator is a pooled estimate of the standard deviation) as a measure of overall bias in that systematic errors of estimation are detected (Yen, 1984); (2) the ratio of standard deviations to assess the degree of homogeneity of variability; and (3) Pearson correlation coefficients to assess the degree of linear association between the two sets of true scores. Pearson correlations were also computed to assess the relation between simulated and estimated thetas.

Results

LOGIST Estimates of Theta vs Simulated Thetas:

The estimated thetas from LOGIST and the simulated thetas are distributed as (Ø,1) over both groups of examinees. While the limits of the theta scale in LOGIST were constrained to be ±3.Ø, the obtained distributions of estimated thetas were truncated at the upper end. The correlations between the estimated and simulated thetas were consistently good over all three conditions (see Table 1).

Insert Table 1 about here.

True score comparisons:

The results for the tests taken, as well as the hypothetical tests not taken are reported in Table 2.

Insert Table 2 about here.

Tests Taken: The overall bias as measured by the standardized mean difference of the true scores is quite small for the tests taken: the bias ranges from approximately zero to 5% of a pooled standard deviation over test forms and conditions. The two sets of true scores are homogeneous with respect to variability; and the correlations are consistently good.

Hypothetical Tests Not Taken: There is a modest indication of overall bias in the standardized mean difference of the true scores for the hypothetical tests not taken. The mean true scores based on LOGIST estimates for test X underestimate the simulated mean true scores in each condition; whereas for test Y, the LOGIST-based mean true scores overestimate the simulated mean true scores. The bias over test forms and conditions ranges from approximately 7% of a pooled standard deviation for test Y in condition III to 17% for test Y in condition II.

The variability is again, quite homogeneous; and the correlations are consistently good over test forms and conditions (see Table 2).

Discussion

The method of simultaneous estimation or concurrent calibration of person and item parameters using LOGIST has been discussed by Pollack (1985), Petersen *et al* (1983), and Hambleton and Swaminathan (1985) in the context of anchor test designs where a common block of items exists and where the comparison groups of examinees may not overlap. Concurrent calibration is used to link items or to equate tests. For this reason, the three conditions reported here involved varying the magnitude of the common block of items from 75% in Condition I, to 50% in Condition II, to no common block (all items unique) in Condition III. However, the similarity of the reported statistics over all three conditions for tests taken, as well as not taken, suggests that the equivalence (within sampling error) of the two groups sampled from the same population overrides the necessity of equating through a common block of items; the "conditions" reported here are better thought of as occasions for replication of results comparing a particular test form's simulated thetas and true scores with the corresponding estimates.

To assess the impact of varying the size of the common block of items where the comparison groups are not equivalent, the input dataset for LOGIST corresponding to the response matrix U would have to be modified for each condition so that the shared items occur only once. Presently, the two tests are adjoined. The consequence of varying the common block size in this study merely introduced varying degrees of test form comparability.

The correlational results were initially surprising inasmuch as varying the common block size or comparability of test forms made no difference: the correlations between simulated and LOGIST-based true scores remained consistently good over all conditions. This was puzzling for it was predicted that the correlations for the hypothetical tests not taken would be especially poor in the third condition (no common block). However, these results may be explained by the fact that the two groups of simulated examinees were representative samples of a normal distribution with mean equal to zero and variance equal to one. A check on the distributions of theta for groups I and II confirmed that while there were minor sample fluctuations, each goup in the simulation phase was representative of the population.

To better understand the impact of equivalent groups in the present study, recall that the number-right true score for a given examinee is a sum of the probabilities (p-values) of responding correctly to each item on a test. A p-value is a function of the estimated theta for that person as well as the estimated item parameters for a particular item. We have, as a consequence of the present design:

$$\zeta_{gk} = f(P_i(\theta_{gk})) \quad \text{and} \qquad (3)$$

$$P_i(\theta_{gk}) = f(\theta_{gk}, a_{gi}, b_{gi}, c_{gi}) \qquad (4)$$

where ζ_{gk} = number-right true score for the k^{th} examinee in the g^{th} group and

$P_i(\theta_{gk})$ = the probability of responding correctly to the i^{th} item given the k^{th} examinee in group g .

In the case where a true score was computed for an examinee, say, in Group I on test X, a test taken for this group and coded as such in a LOGIST run, we have p-values of the form:

$$P_i(\theta_{Ik}) = f(\theta_{Ik}, a_{Ii}, b_{Ii}, c_{Ii}) \quad . \qquad (5)$$

Note that the p-value in equation (5) is a function of item parameters calibrated on the same group from which the theta estimate is derived. However, in the case where a true score was computed for an examinee, say, in Group I on Test Y , a hypothetical test not taken for this group, we have p-values of the form:

$$P_i(\theta_{Ik}) = f(\theta_{Ik}, a_{IIi}, b_{IIi}, c_{IIi}) \quad . \qquad (6)$$

In equation (6) we see that the item parameters are not calibrated on the same group from which the examinee came.

If the items are calibrated on a group of examinees that is not quite equivalent to the comparison group, and if the item parameters are similarly distributed over both tests , then the estimated mean true score difference where one of the groups has not actually taken the test will probably be a somewhat biased estimate of the mean group difference had both groups taken the test. To the extent that groups of examinees are equivalent and representative samples of some population, and the items of the tests are sampled from the same unidimensional domain, bias should be eliminated since estimates of theta are invariant with respect to the particular subset of items answered; similarly, estimates of item parameters are invariant to the particular subgroup or sample of examinees. In other words, as long as the complete latent space is correctly specified, bias should be effectively zero.

The results reported in Table 2 suggest that the bias is quite small for tests taken since examinee true scores were based on items calibrated on the same group from which examinees came. We may infer that the LOGIST estimates of b's and thetas were quite faithful (within a scale transformation) to the simulated parameters when a's and c's were fixed. Note also that the samples of the present study were very small: 100 simulated examinees for each of two tests, and 20 items per test.

In conclusion, the modest overall bias reported here for what could be considered a "best case" scenario suggests that until the effects of using non-equivalent samples , and/or using sets of item parameters which are discrepant over the tests of interest are known, the comparison of cohorts in the manner described above should remain experimental. In particular, the method should not be used for assessing cohort differences until the likely direction and magnitude of bias given the constraints of any specific study are well understood.

Figure 1

Partitioned Matrix U of Simulated

Examinees by Item Responses

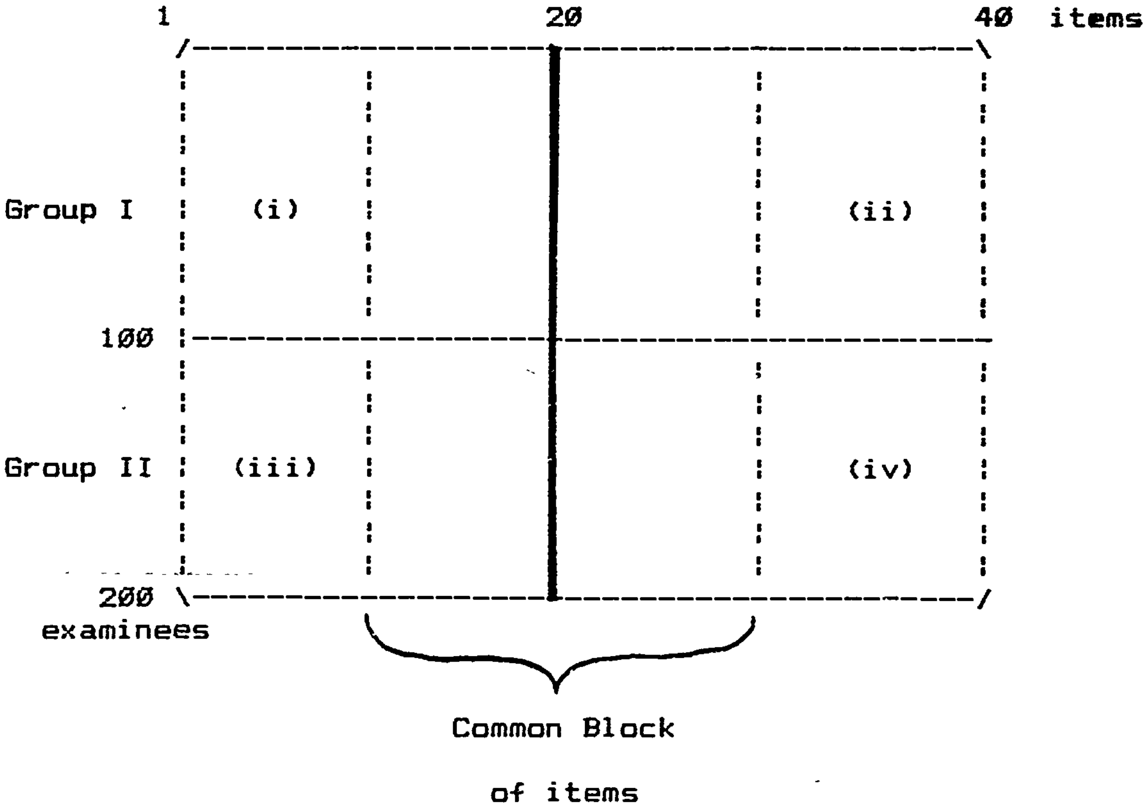


Figure 2

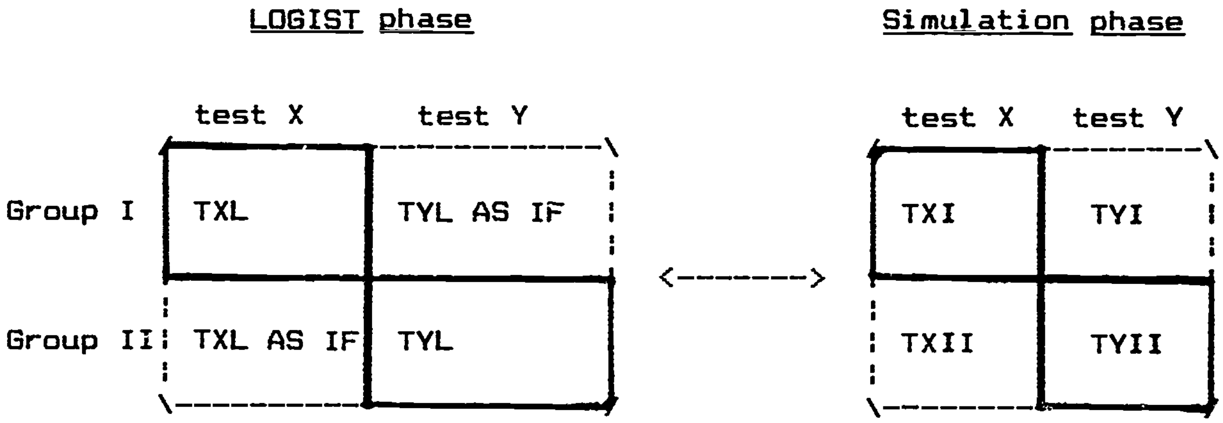
Design of a LOGIST Input Dataset

```
                      test X                   test Y

             1                     2Ø                     4Ø items
             /----------------------------------------------\
             :                     :                      :
             :                     :                      :
             :                     :                      :
Group I      :    1's and Ø's      :         3's          :
             :                     :                      :
             :                     :                      :
         1ØØ :---------------------:----------------------:
             :                     :                      :
             :                     :                      :
Group II     :        3's          :      1's and Ø's     :
             :                     :                      :
             :                     :                      :
         2ØØ \----------------------------------------------/
     examinees
```

Figure 3

Corresponding Blocks of True Scores for Tests Taken and for Tests Not Taken Based on LOGIST Estimates and on Simulated Values



LOGIST phase

	test X	test Y
Group I	TXL	TYL AS IF
Group II	TXL AS IF	TYL

Simulation phase

	test X	test Y
Group I	TXI	TYI
Group II	TXII	TYII

where ▬ tests taken

_____ tests NOT taken

TXL = true scores on test X, Group I
TXL AS IF = hypothetical true scores on test X, Group II
TYL = true scores on test Y, Group II
TYL AS IF = hypothetical true scores on test Y, Group I

/ Scores are functions of LOGIST estimates

TXI = true scores on test X for Group I
TXII = true scores on test X for Group II
TYI = true scores on test Y for Group I
TYII = true scores on test Y for Group II

/ Scores are functions of simulation parameter values

Table 1

LOGIST Estimates of Theta *vs* Simulated Thetas

Condition	r(θ_{LOGIST}, θ_{simuln})	Limits of simulated thetas	Limits of LOGIST estimates of theta
I	.894	-2.59, 2.60	-3.00, 1.68
II	.881	-2.48, 2.54	-3.00, 1.84
III	.868	-2.48, 2.76	-3.00, 1.74

Note: The missing LOGIST estimates for examinees with perfect scores were later coded as 1.8 in Conditions I and III, and 1.9 in Condition II . These values were used in the computation of the correlations reported here.

Table 2

True Score Comparisons

Tests Taken

Condition	I	II	III
Relative size of common block of items	75%	50%	none
$(\overline{TX}-\overline{TXL})/SD_p$	.002	-.033	.026
$(\overline{TY}-\overline{TYL})/SD_p$	-.024	-.010	.054
SD_{TX}/SD_{TXL}	.934	1.032	.951
SD_{TY}/SD_{TYL}	.953	.972	.886
r (TX,TXL)	.933	.910	.913
r (TY,TYL)	.903	.905	.904

Tests NOT Taken

Condition	I	II	III
$(\overline{TX}-\overline{TXL})/SD_p$	.127	.118	.152
$(\overline{TY}-\overline{TYL})/SD_p$	-.128	-.171	-.067
SD_{TX}/SD_{TXL}	.941	.974	.889
SD_{TY}/SD_{TYL}	.936	1.031	.954
r (TX,TXL)	.904	.904	.902
r (TY,TYL)	.931	.910	.914

Note: TX = simulated true scores on test X
TY = simulated true scores on test Y
TXL = LOGIST-based true scores on test X
TYL = LOGIST-based true scores on test Y
SD_p = pooled estimate of the standard deviation
r = Pearson correlation coefficient

References

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. New York: Erlbaum.

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

Petersen. N. S., Cook, L. L., & Stocking, M. L. (1983). *IRT versus conventional equating methods: A comparative study of scale stability*. Journal of Educational Statistics, *8*(2), 137-156.

Pollack, J. (1985, March). *Using IRT methods to put NLS and HSB tests on a common scale*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, Illinois.

SAS Institue Inc. (1982). *SAS user's guide: 1982 edition*. Cary, NC: SAS Institute Inc.

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). *LOGIST user's guide: LOGIST 5, version 1.0*. Princeton, NJ: Educational Testing Service.

Yen, W. M. (1984). Obtaining maximum likelihood trait estimates from number-correct scores for the three-parameter logistic model. *Journal of Educational Measurement*, *21*(2), 93-111.