DOCUMENT RESUME

ED 246 079 TM 840 350

AUTHOR	Chang, S. Tai; Bashaw, W. L.
TITLE	Characteristics of Anchor Tests.
PUB DATE	Apr 84
NOTE	38p.; Paper presented at the Annual Meeting of the American Educational Research Association (68th, New Orleans, LA, April 23-27, 1984).
PUB TYPE	Speeches/Conference Papers (150) -- Reports - Research/Technical (143)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Aptitude Tests; *Difficulty Level; *Equated Scores; Junior High Schools; *Latent Trait Theory; *Test Items; *Test Length; Test Norms
IDENTIFIERS	*Anchor Tests; Item Calibration; Otis Lennon School Ability Test; Person Fit Measures; Rasch Model; *Test Linking

ABSTRACT

The purpose of this study was twofold: to investigate to what extent characteristics of anchor tests may affect precision of item calibration, and to estimate to what extent precision of item calibration may be affected by removal of persons whose response patterns deviate from those normally expected from the Rasch one-parameter logistic model. Three characteristics of anchor tests were under consideration: the number of anchor items, and the range and average of difficulties of the anchor items. The data were taken from the nation-wide norming data of the Otis-Lennon School Ability Test, Form R, Intermediate Level. Anchor test characteristics did not show systematic effects on final calibration results. The removal of misfitting persons was detrimental to calibration results. Further studies are needed to clarify the effects of anchor test designs and person fit on linkings. (BW)

Characteristics of Anchor Tests

S. Tai Chang and W. L. Bashaw

University of Georgia

Presented at the 1984 Annual Meeting of the American Educational Research Association, New Orleans, La., April, 1984. The authors express their gratitude to the Psychological Corporation for making available to us the 1977 norming data of the Otis-Lennon School Ability Test. We are also grateful to R. Robert Rentz for providing computer programs and consultation.

INTRODUCTION

For decades, psychometricians have been striving to enhance objectivity, accuracy, and efficiency of mental measurement. The most important developments in recent years are probably the latent trait theory and its applications, particularly, computerized tailored testing. Availability of a large homogeneous item pool (of 200 or more items) is usually prerequisite for tailored testing and/or some other more advanced applications of the latent trait theory. Also required is that all the items in the pool be precisely calibrated on a single common scale. Since it is not feasible to administer a very long test to any single group of examinees, items for a pool are usually collected from several item sets which are calibrated on independent groups. (There are also some other factors that dictate collection of items from multiple item sets on independent groups, for instance, for updating an item pool or for constructing tests with comparability across time.) Unless equivalent groups are used in calibration, some conversion is usually needed to link item parameter estimates obtained from separate groups.

Linking two sets of separately calibrated items can be accomplished through either a common group of examinees who take both tests or through a set of common items (known as an anchor test) taken by different groups of examinees. Since usually it is inconvenient to use a common group of persons for linking, the anchor test approach becomes the primary means for linking and is the concern of the present study.

There are many psychometrical models subsumed to the latent trait theory. One of the more popular model is Rasch's one parameter logistic (1PL) model. Review of psychometric literature reveals that although 1PL model has attracted a large number of equating studies, only a few linking studies have been conducted. Equating and linking are symmetrical procedures and have some similarities; nevertheless, there are also important differences. Most of all, equating deals directly with accuracy of measured ability scores, whereas linking deals directly with precision of calibrated item estimates which eventually affect accuracy of measured ability scores.

In some equating studies, effect of length of anchor tests on accuracy of equated score was investigated, but no consistent results were reached. This is primarily due to lack of a good criterion for evaluation of equated scores. If the Monte Carlo method is used, the criterion problem is solved but the results may not conform to reality. A better solution is to employ several test forms to constitute a circular chain and through consecutive equating the initial test form will be finally equated to itself. Consistency of ability scores then becomes an evaluation criterion. Since it is extremely laborious, this approach is seldom used by researchers.

As far as linking is concerned, current knowledge about anchor test length and other characteristics is limited. From their item calibration experience using the classical test model, McBride and Weiss (1974) claimed that 40 to 60 anchor items may be needed to calibrate an item pool. Based on theoretical values of standard errors of item estimates, Wright considers a sample size of 400 persons and an anchor test of 10 to 20 items as sufficient for most linking situations. Wright contends that ten anchor items may be adequate if the items are good (Wright, 1977).

While most linking studies dealt conceptually with linking problems, one empirical study (McKinley & Reckase, 1981) investigated effects of sample size and anchor test length on precision of item parameter estimates. There were three levels in test length-- 5, 15, and 25 items. Correlation between linked estimates and estimates obtained from the original total sample was used as an evaluation criterion. Obtained correlation values under all conditions were close to unity. Despite trivial differences among the correlations , results generally indicated the longer the anchor test and the larger the sample size, the better the precision. Only in one condition was the five-item anchor better than the fifteen-item anchor. The investigators thus thought a five-item anchor might be adequate, but a fifteen-item anchor was suggested.

However, the correlations used to evaluate calibration and linking results may be affected by distributions of item parameter estimates and does not necessarily reflect magnitude of errors introduced through estimation and linking processes. Moreover, factors other than size of samples and length of anchor tests also need to be identified and investigated to provide guidelines for construction of anchor tests for linking and guarantee that desired precision of item calibration can be reached.

In item calibration, misfit of an item to the Rasch model can be due to aberrant test-taking behavior of a few persons just as it can be due to a general flaw in the item itself. It is conjectured that impact of irregular person responses can be serious when examinee sample size is small and that even large samples may not obliterate contaminating influence of irregular person responses (Wright & Stone, 1979, p. 82). Few previous studies deal with person fit problem and no one has investigated how removal of misfitting persons affects calibration of item pools.

The purpose of this study was twofold. One was to investigate to what extent characteristics of anchor tests may affect precision of item calibration. The other purpose was to estimate to what extent precision of item calibration may be affected by removal of persons whose response patterns deviate from what are normally expected from 1PL model. Three characteristics of anchor tests were under consideration, namely, test length, test width, and test height. These three characteristics correspond, respectively, to number of anchor items, range and average of difficulties of the anchor items (Wright & Stone, 1979, p. 133)

METHODS AND PROCEDURES

Design of the Study

The fundamentals of the design of this study can be perceived as similar to Angoff's (1971) Equating Design IV or Equating Design VI. The essentials of these designs are as follows:

Test (Form) X is administered to Group A; Test (Form) Y is administered to Group B. Tests (Forms) X and Y have a set of items in common (i.e., an anchor test). The anchor test is administered to both Group A and B and is used to adjust differences that exist between the two tests (forms).

In the present study, a nonanchor test was treated as if it were two different tests (i.e., analog of equating a test to itself). This kind of treatment was first used in equating research by Levine (1955). A number of more recent equating studies also used it (e.g., Green, 1980; Marco, Petersen, & Stewart, 1979; Pettie, 1981).

The present research was conducted in a fashion of an ex post facto experiment. Nonanchor test items calibrated with groups at two different grade levels were linked onto a base metric through 22 different anchor tests. The anchor tests differed from one another primarily in test height, width, and/or length. At the two grade levels, a pair of random samples of 1000 examinees each was drawn from a data base. The examinee sampling was replicated three times without replacement. Calibrating and linking with each pair of samples were performed under two different situations. In one situation, misfitting persons were not screened and the intact samples were used. In the other situation, misfitting persons were detected and excluded from the groups. Since there were two situations, three replications of sample pairs, and 22 anchor tests, the total number of calibrations is 2 * 3 * 2 * 22 = 264. Final results of item calibration were evaluated in term of fidelity of item estimates.

Data Base

The data base used in the present study was taken from the nation-wide norming data of the Otis-Lennon School Ability Test(OLSAT), Form R, Intermediate level. The Intermediate level was designed for students in grades 6, 7, and 8. There are 80 items at this level in the Form R of the Otis-Lennon test. This form and level of the test was normed in fall, 1977. The sixth and the seventh graders' responses on the items were used in this research as the data base. There are responses from 11,776 sixth grade students and 11,020 seventh grade students in the data base (Otis & Lennon, 1979).

Preliminary Analysis

The sixth and the seventh grade students in the data base were treated as one population. All the items in the data base were calibrated with this population twice, once with the intact population and once with data of non-fitting persons deleted. These analyses yielded information regarding difficulty (DIFF), standard error of difficulty (STDERR), mean square fit (MSFIT), and slope of item response characteristic curve (i.e., discrimination) for each item. Table 1 and Table 2 show the item difficulty and related information from these calibrations. The item difficulty values and the metric thus defined, as shown in Table 1, were taken as item parameters and as the base metric for subsequent analyses when all sample data were used. Table 2 shows similar information for subsequent analyses from which misfitting persons were removed from samples.

The population was then separated according to subjects' grade levels into two subpopulations. All the items in the data base were calibrated with each of the two subpopulations regardless of misfitting persons. Item difficulty values obtained from these two calibrations were plotted against each other to screen of possible outliers. An outlier was loosely defined as ,on the plot, a point that obviously deviated from the best fitted straight line of unit slope. The reason to screen and eliminate misfitting items with this particular method rather than employing slope and/or mean square fit values was that misfitting items could be better judged from such direct fitting results than from some indicators (Cf. Rentz, 1975).. Slope and mean square fit values indicate item misfit in term of extent rather than type. Both of them lack definite criterion values for identifying misfitting items. Figure 1 presents the screening plot. No item was seen as an outlier; thus, all the 80 items in the data base were retained.

Construction of Anchor Tests and Nonanchor Tests

Construction of anchor tests and nonanchor tests was based on item parameters obtained from the calibration with the total population without excluding misfitting persons. Items were separated into two sets. One set contained 50 items and served as an item pool for constructing anchor tests; the other set contained the remaining 30 items and was equally devided into three subsets for constructing nonanchor tests.

Anchor items were selected according to difficulty parameter values, such that difficulty values for 40 items spanned the range of - 2 to + 2 logits and were approximately equally spaced in that interval. Another 10 items were chosen approximately equally spread in the range of - 0.5 to 0.5 logits.

Twenty two anchor tests were constructed. One anchor test comprised all the anchor items. A second and a third anchor test consisted, respectively, of the five and the ten best fitting anchor items, best fitting in the sense of having mean square fit values and slope index values near unity. The other 19 anchor tests differed from one another in the design of test height, width, and/or length. Test height, width, and length are synonyms, respectively, for average item difficulty, range of item diffculty, and number of items in a test. The 19 designed anchor tests centered around one of the following three height levels: - 1.0, 0.0, and 1.0 logits. At the - 1.0 and 1.0 height levels, there were 1.0 and 2.0 width levels. At the 1.0 width level, there were two length levels--lengths of five or ten items. At the 2.0 width level, there were three length levels--five, ten, and twenty items. At the 0.0 height level, there were three width levels-- 1.0, 2.0 and 3.0. At each of these three width levels, test lengths were five, ten, or twenty items. Table 3 lists these specifications of the anchor tests and the actual height of each constructed anchor test.

Two nonanchor tests were constructed from the 30 items reserved for this purpose. One subtest of ten items was used in both nonanchor tests. These common items were combined with one of the other two ten-item subsets to form nonanchor tests for the 6th and 7th graders, respectively.

Table 4 lists the items assembled into each anchor test as well as into nonanchor tests. The numerals in the table are the same as actual item numbers in the Otis-Lennon School Ability Test.

Examinee Samples and Calibration Procedures

A pair of random samples of 1000 examinees each was drawn from the sixth and the seventh grade subpopulations. This examinee sampling process was replicated three times, without replacement, resulting in three different sample pairs. Sampling examinees from the subpopulations was performed in two stages. In the first stage, random numbers were generated using a uniform random number function and the numbers were attached to examinees' data records. For each of the two subpopulations, three independent samples, each of a size slightly over 1000, were then produced by specifying three mutually exclusive ranges of the random numbers. In the second stage, exactly 1000 examinees' records were randomly taken from each sample. A computer software system, SAS, was utilized to accomplish the sampling of examinees (Cf. Ray, 1982; Council, 1980, p. 152).

Items in each of the two nonanchor tests, along with items in each of the 22 anchor tests, were calibrated with respective examinee samples in each sample pair. Nonanchor test items calibrated with lower grade level samples as well as the nonanchor items calibrated with higher grade level samples were then placed, through different anchor tests, onto the base metric. The linked item estimates obtained from environments of different anchor tests and from each sample were compared with their item parameter. This was done for both intact-sample and excluding-misfitting-sample situations.

Exclusion of Misfitting Persons

Statistical procedures for identifying misfitting persons are illustrated in Wright and Stone (1979). According to them (p. 168), the person-fit statistic is more or less normally distributed but with wider tails. They consider a rejection level of about 2.0 as conservative and 3.0 acceptable. The present study used t=2.5 as the critical value to detect and exclude misfitting persons.

Translation of Item Estimates

Conversion is needed to place item estimates from different data sets onto a common metric. Rationale and method for translating Rasch item difficulty estimates and/or examinee ability estimates from one test scale to another test scale have been described by Rentz and Bashaw (1975, 1977). More illustrations of linking together two sets of item estimates through an anchor test can be found in Mead (1981), Kreines and Mead (1979), and Wright and Stone (1979). With the Rasch model, whenever two separately calibrated tests both measure the same trait and both fit the model, the test scale defined for them will have the same units, but different origins (Rentz & Bashaw, 1977,

p. 162). To link together the two sets of items, usually, the scale of one of the tests is chosen to serve as a base metric and the other scale is to be adjusted. Adjustment is performed simply by finding the difference (in log units) between the average of anchor item difficulties on the base metric and the average of anchor item difficulties on the test scale being adjusted. The difference is then used as an additive constant to translate the nonanchor item estimates from the scale being adjusted onto the base metric. Linking two sets of item estimates to a "preexisting" base metric is just the same as linking together two tests, except two, rather than one, sets of estimates need to be translated onto the particular base metric.

Evaluation Method and Criterion

Final results of item calibration (i.e., linked item difficulty estimates) were evaluated in term of fidelity of item estimates with an absolute criterion. Fidelity deals with discrepancies between each set of linked item estimates and the parameters. The criterion for evaluation was distance from an absolute value of zero. Distance is used here for values of fidelity discrepancy, irrespective of arithmetic signs.

Theoretically, discrepancy values should be close to zero if item calibration and linking can be done perfectly. Although perfect calibration actually never can be expected, distance from a value of zero can provide some useful information with regard to evaluating final calibration results. Descriptive statistics such as mean, standard deviation, minimum, and maximum of the distance values were calculated. In addition, means of discrepancy values were also computed to indicate direction of possible bias.

Data Analysis

Estimations of item difficulty and related information were done by using a computer program called TRIAN; more exactly, June 1985 version of the TRIAN program was employed. The program was prepared by R. Robert Rentz to provide both traditional item analysis statistics and Rasch model analyses. The TRIAN program can perform person-fit analysis while estimating item and ability parameters. Unconditional maximum likelihood procedures of Wright and Panchapakesan (1969) and Wright and Mead (1976) have been adapted in the TRIAN program (Note 1). Programs written with SAS were used in computing translation constants, in linking difficulty estimates of the nonanchor test items to the base metric, and in performing the evaluation.

RESULTS

Sample Sizes

In Rasch calibration responses from any person who answers no item correctly (a zero test score) or all items correctly (a perfect test score) are not used. In the present study, total number of persons removed from any sample due to zero or perfect score in any calibration never exceeded seven. In the calibrations with removal of misfitting persons, the numbers of misfitting persons removed from samples varied from 34 to 132, but most were between 50 to 90.

Evaluation of Fidelity Discrepancy

An examination of means of discrepancy values does not reveal any systematic bias in the linked item estimates resulted from anchor test characteristics or person fit situations.

Evaluation of Fidelity Distance

Results of evaluation on fidelity distance are shown in Table 5 to Table 8. In these tables, characteristics of the anchor tests are the same as those in Table 3. Comparisons of the statistics across tests and samples for each grade and each situation do not show any systematic effect of anchor test characteristics on the final calibration results. Comparisons across situations for each grade seem to indicate that removal of misfitting persons makes calibration results slightly worse, if there are any differences, than no removal.

DISCUSSION

The fact that anchor test characteristics did not show systematic effects on final calibration results was considered to be probably due to one or both of the following two reasons:

A. The OLSAT data fit the Rasch model very well, and/or

B. There exists an inherent limitation on the precision of calibration that can be attained with the method used in the present study.

Theoretically, when data fit the model perfectly, linking results should be the same regardless of any difference in the characteristics of anchor tests. In other words, it should not matter what items constitute an anchor test if data fit the model perfectly.

The computer program TRIAN produces two item fit statistics: index of mean square fit and index of slope values. Overall fit of a set of data to the model can be evaluated through the values of these indices. However, due to the algorithm in the program, the values obtained for the index of slope from each calibration are very rough indicators of item fit. The mean square fit value for each item is a more accurate index but not a perfect one. The values of mean square fit for the OLSAT items indicate a good overall fit of the OLSAT data to the Rasch model.

An alternative approach to evaluate fit of data to the model is by examining plots of item difficulty values obtained from two subgroups of examinees. If there is a good fit then the points in a plot should fall along a 45 degree angle line. Again, this approach is merely a rough way of evaluation of fit. The fit of the OLSAT data to the Rasch model can be considered excellent if

We evaluate overall fit of the data to the model by way of the plot shown in Figure 1.

Through calibration and linking procedures, some errors are inevitably introduced into item parameter estimates. It has long been known that random sampling does not necessarily generate equivalent groups. From the replications of the samples in the present study, it was found that sample fluctuation may cause substantial errors even when sample size is as large as 1000. Also noticed was that sometimes estimates on individual items also fluctuate. The linking method used in this study makes adjustments on estimates for nonanchor items in each design circumstance by merely a single additive constant. While this simple method very accurately adjusted estimates for most items in most cases, it was not unexpected that it would have failed to perform well on some items in some cases. Results from an evaluation of fidelity discrepancy on individual items seemed to indicate that such failures were random events. Since the errors from these major sources were random and inevitable, systematic error brought about by anchor test characteristics, if any , may not be easily detectable. It should be noted that with empirical data it is very difficult to isolate each source of error. But for all practical purposes, the concern is usually the magnitude of total error associated with each item rather than the distinction of different error sources.

The fact that removal of misfitting persons seems detrimental to calibration results was attributed to one or both of the following two possible reasons:

A. Expansion in ranges of item estimates and scales defined by the estimates, and/or

B. Appropriateness of the person fit statistics.

In the present study it was found that item estimates and scales defined by the estimates were "stretched out" to about three tenths (in logit units) on both easy and hard ends when misfitting persons were excluded from calibrations. An examination of removed persons reveals that most misfitting persons were at the low ability end, but there were also some misfitting persons at the high ability end and some others in the middle. It is our conjecture that expansions in the item estimates (equivalently in the scales) allow somewhat larger errors to be introduced into the estimates.

Whether person fit statistics t is an appropriate measure for identification of misfitting persons is a question for which we do not have a ready answer at the present time. The person fit t index is a summary indicator of misfitting responses for each person. It is easy to use, but from a glance at some items misanswered by a number of misfitting persons, we suspect that the person fit t index may not be a valid way to screen irregular response patterns.

CONCLUSIONS AND SUGGESTIONS FOR FURTHER STUDIES

Based on the findings from this study, it seems two temporary conclusions can be drawn:

1. Linking can probably be done quite effectively over a wide range of anchor test designs, and

2. Removal of misfitting persons using person fit t index may work detrimentally.

Some further studies are certainly needed to better clarify the effects of anchor test designs and person fit on linkings. The OLSAT test used in this study is an unambiguously unidimensional test. The data seem to fit the model very well. For most achievement tests, dimensionality may not be so unambiguous and model-data fit may not be very good. In such cases whether the findings from this study can still hold needs to be investigated. After all, achievement testing is the area to which latent trait models are most likely to be applied. Whether there exists an inherent limitation on the precision that can be attained with the linking method used in this study can be investigated by applying the method to a variety of larger sample sizes. If errors obtained with some larger sample sizes approach nearly the same magnitude , a clear limitation can then be concluded. Different measures of person fit and their effects on linking also need to be more thoroughly studied before we can firmly declare who are misfitting persons and whether they should or should not be removed from calibration.

Reference Note

1. Rentz, R. R. (1983). TRIAN item analysis: Documentation for FORTRAN computer program (Version of 11 June 1983).

References

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.) Educational measurement (2nd ed.) Washington, D.C.: American Council on Education.

Council, K. A. (Ed.). (1980). SAS applications guide, 1980 edition. Cary, N.C.: SAS Institute Inc.

Green, J. C. (1985). An investigation of two linear equating methods where abilities vary. Unpublished doctoral dissertation, Florida State University.

Kreines, D. C. & Mead, R. J. (1979). Equating tests with the Rasch model. Paper presented at the annual meeting of the National Council on Measurement in Education.

Levine, R. S. (1955). Equating the score scales of alternate forms administered to samples of different ability (RB-55-23). Princeton, N.J.: Educational Testing Service.

Marco, G. L., Petersen, N. S., & Stewart, E. E. (1979). A test of the adequacy of curvilinear score equating models. Paper presented at the 1979 Computer Adaptive Testing Conference, Minneapolis.

McBride, J. R., & Weiss, D. J. (1974). A word knowledge item pool for adaptive ability measurement (Research Report, 74-2). University of Minnesota, Psychometric Methods Program.

McKinley, R. L., & Reckase, M. D. (1981). A comparison of procedures for constructing large item pools. Columbia, Mo.: Missori University, Tailored Testing Research Laboratory.

Mead, R. J. (1981). Basic ideas in item banking. Paper presented at the annual meeting of the National Council on Measurement in Education.

Otis, A. S., & Lennon, R. T. (1979). Otis-Lennon School Ability Test. Harcourt Brace Jovanovich, Inc.

Pettie, A. L. (1981). Rasch model anchor test equating and person fit: An examination of achievement test equating and person fit. Unpublished doctoral dissertation, the Florida State University.

Ray, A. A. (Ed.). (1980). SAS user's guide: Basics, 1982 edition. Cary, N.C.: SAS Institute Inc.

Rentz, C. C. (1975). An investigation of the invariance properties of the Rasch model parameter estimates. Unplished doctoral dissertation, University of Georgia.

Rentz, R. R., & Bashaw, W. L. (1975). Equating reading tests with the Rasch model (Vols 1 & 2). Athens, Ga.: University of Georgia, Educational Research Laboratory.

Rentz, R. R., & Bashaw, W. L. (1977). The national reference scale for reading: An application of the Rasch model. Journal of Educational Measurement, 14, 161-179.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. Journal of Educational Measurement, 14, 97-116.

Wright, B. D., & Mead, R. J. (1976). BICAL: Calibrating items with the Rasch model (Research Memorandum No 23) University of Chicago, Department of Education, Statistical Laboratory.

Wright, B. D., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. Educational and Psychological Measurement, 29, 23-48.

Wright, B. D., & Stone, M. H. (1979). Best test design: A handbook for Rasch measurement. Chicago: MESA Press.

FIGURE 1: PLOT OF ITEM DIFFICULTIES OBTAINED FROM 6TH GRADERS AGAINST ITEM DIFFICULTIES OBTAINED FROM 7TH GRADERS
LEGEND: A= 1 ITEM, B= 2 ITEMS, AND C= 3 ITEMS

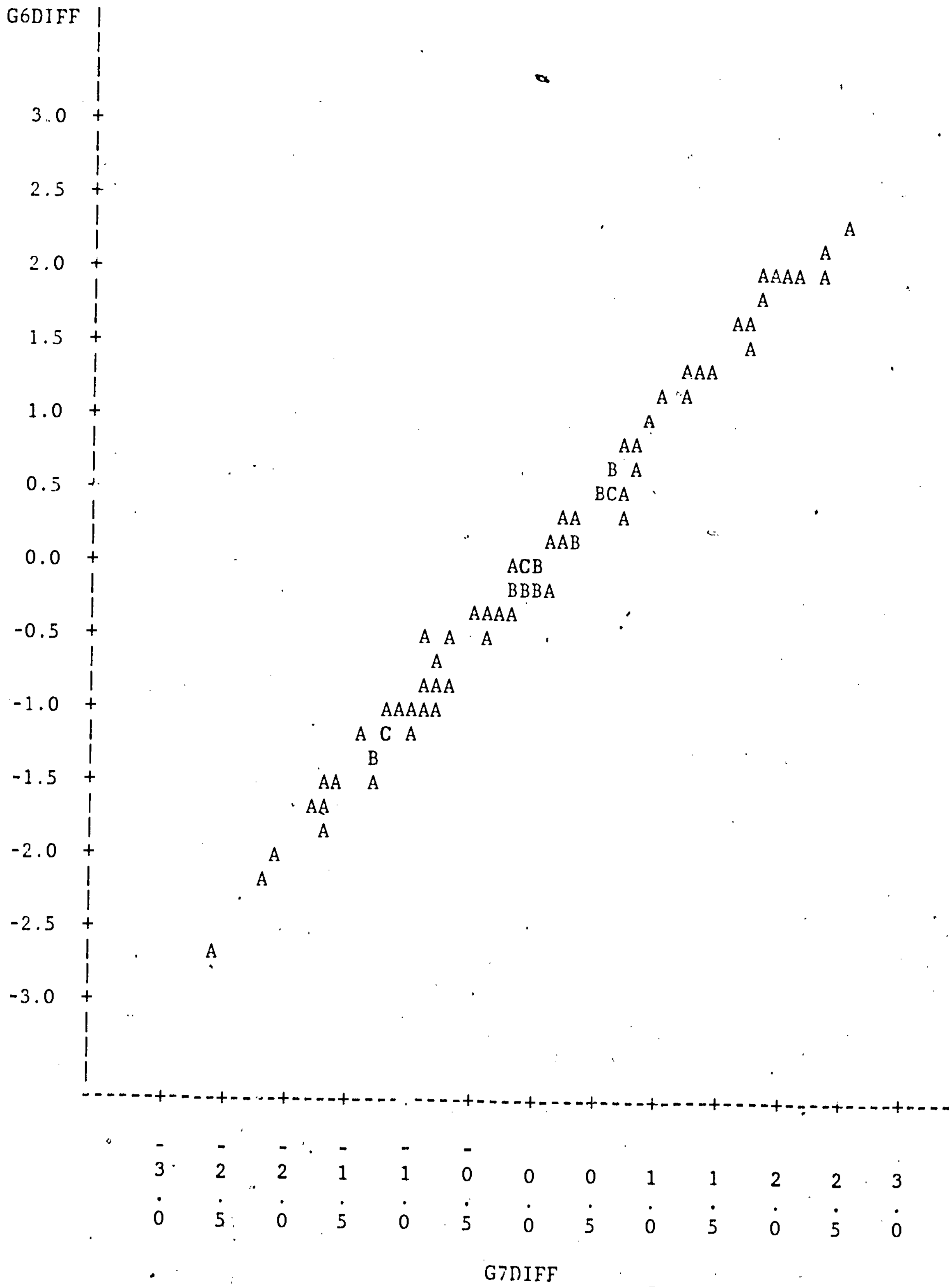

TABLE 1
RASCH CALIBRATION OF CLSAT USING TOTAL POPULATION

ITEM	DIFF	STDERR	MSFIT	SLOPE
1	-2.593	0.023	1.04	0.67
2	-1.698	0.018	0.80	1.41
3	-1.111	0.016	0.80	1.52
4	-1.554	0.017	0.98	0.93
5	-1.765	0.018	1.18	0.61
6	-2.127	0.020	0.71	1.40
7	-1.211	0.016	1.01	0.89
8	-0.894	0.016	0.85	1.42
9	-1.428	0.017	1.10	0.69
10	-0.931	0.016	1.01	0.86
11	-1.263	0.016	0.88	1.24
12	-2.019	0.019	0.65	1.69
13	-1.606	0.018	0.71	1.66
14	-0.568	0.015	0.88	1.31
15	-1.211	0.016	0.83	1.41
16	-1.298	0.017	0.90	1.13
17	-0.773	0.015	0.91	1.23
18	-1.684	0.018	1.04	0.79
19	-1.103	0.016	1.05	0.80
20	-0.425	0.015	0.88	1.33
21	-0.701	0.015	0.92	1.15
22	-1.096	0.016	0.78	1.62
23	-0.746	0.015	0.91	1.21
24	0.241	0.015	1.06	0.81
25	-1.361	0.017	0.90	1.11
26	-0.112	0.015	0.98	1.03
27	-1.172	0.016	0.89	1.23
28	-0.007	0.015	1.01	0.90
29	-0.140	0.015	0.94	1.15
30	-0.416	0.015	1.03	0.91
31	-0.408	0.015	1.06	0.79
32	-0.092	0.015	0.87	1.38
33	-0.729	0.015	0.90	1.27
34	-0.003	0.015	0.93	1.19
35	-0.066	0.015	0.98	1.05
36	-0.980	0.016	0.90	1.18
37	-0.275	0.015	0.95	1.17
38	0.533	0.015	1.03	0.89
39	-0.025	0.015	1.13	0.61
40	0.118	0.015	0.99	1.03

TABLE 1 cont.

ITEM	DIFF	STDERR	MSFIT	SLOPE
41	0.625	0.015	1.07	0.79
42	-0.992	0.016	0.91	1.21
43	0.530	0.015	1.19	0.50
44	-0.050	0.015	0.96	1.08
45	-0.122	0.015	1.08	0.68
46	0.499	0.015	0.91	1.23
47	-0.049	0.015	0.96	1.12
48	-0.004	0.015	0.93	1.20
49	-0.243	0.015	1.10	0.65
50	0.491	0.015	1.23	0.37
51	0.179	0.015	0.98	1.06
52	-0.191	0.015	0.90	1.29
53	0.572	0.015	1.13	0.61
54	0.704	0.016	0.84	1.47
55	0.236	0.015	1.00	0.98
56	0.844	0.016	1.03	0.87
57	1.028	0.016	1.03	0.89
58	0.023	0.015	0.95	1.13
59	0.937	0.016	0.98	1.04
60	0.311	0.015	0.91	1.33
61	0.783	0.016	1.05	0.84
62	0.601	0.015	0.94	1.14
63	1.194	0.017	1.16	0.67
64	1.626	0.018	1.17	0.72
65	1.365	0.017	1.08	0.72
66	0.259	0.015	0.95	1.11
67	1.302	0.017	0.97	0.95
68	0.555	0.015	0.91	1.23
69	0.621	0.015	1.11	0.62
70	1.202	0.017	1.04	0.87
71	1.864	0.019	1.23	0.51
72	1.856	0.019	0.85	1.19
73	2.434	0.022	1.08	0.53
74	1.976	0.020	1.21	0.47
75	2.127	0.020	1.45	0.25
76	1.599	0.018	1.21	0.54
77	1.716	0.018	1.42	0.20
78	2.223	0.021	1.42	0.31
79	2.028	0.020	1.24	0.53
80	2.042	0.020	1.46	0.33

TABLE 2
RASCH CALIBRATION OF OLSAT USING TOTAL
POPULATION EXCLUDING MISFITTING PERSONS

ITEM	DIFF	STDERR	MSFIT	SLOPE
1	-2.902	0.029	1.06	0.72
2	-1.845	0.021	0.82	1.42
3	-1.268	0.018	0.83	1.44
4	-1.715	0.020	1.07	0.75
5	-1.850	0.021	1.27	0.56
6	-2.456	0.025	0.77	1.18
7	-1.299	0.018	1.06	0.81
8	-0.996	0.017	0.87	1.40
9	-1.489	0.019	1.16	0.64
10	-1.028	0.017	1.05	0.75
11	-1.376	0.019	0.91	1.16
12	-2.367	0.024	0.67	1.58
13	-1.861	0.021	0.73	1.56
14	-0.642	0.016	0.91	1.27
15	-1.364	0.018	0.85	1.30
16	-1.473	0.019	0.96	1.01
17	-0.854	0.017	0.93	1.17
18	-1.813	0.021	1.14	0.72
19	-1.198	0.018	1.12	0.69
20	-0.481	0.016	0.90	1.27
21	-0.775	0.017	0.95	1.11
22	-1.248	0.018	0.79	1.56
23	-0.855	0.017	0.94	1.10
24	0.245	0.016	1.07	0.75
25	-1.558	0.019	0.97	0.91
26	-0.134	0.016	1.00	0.99
27	-1.317	0.018	0.94	1.15
28	0.024	0.016	1.00	0.92
29	-0.158	0.016	0.95	1.13
30	-0.427	0.016	1.04	0.88
31	-0.428	0.016	1.08	0.69
32	-0.127	0.016	0.90	1.34
33	-0.819	0.017	0.92	1.23
34	-0.012	0.016	0.95	1.16
35	-0.037	0.016	0.97	1.11
36	-1.104	0.018	0.94	1.04
37	-0.303	0.016	0.96	1.12
38	0.552	0.016	1.03	0.85
39	-0.005	0.016	1.12	0.56
40	0.120	0.016	1.00	0.98

TABLE 2 cont.

ITEM	DIFF	STDERR	MSFIT	SLOPE
41	0.669	0.016	1.05	0.79
42	-1.138	0.018	0.96	1.06
43	0.582	0.016	1.16	0.50
44	-0.053	0.016	0.97	1.06
45	-0.135	0.016	1.10	0.58
46	0.531	0.016	0.92	1.26
47	-0.061	0.016	0.97	1.09
48	-0.004	0.016	0.94	1.20
49	-0.244	0.016	1.12	0.59
50	0.598	0.016	1.17	0.41
51	0.178	0.016	0.99	1.01
52	-0.226	0.016	0.92	1.24
53	0.662	0.016	1.10	0.65
54	0.714	0.016	0.87	1.46
55	0.266	0.016	1.00	1.00
56	0.923	0.017	1.00	0.93
57	1.139	0.017	0.97	1.00
58	0.026	0.016	0.96	1.13
59	1.010	0.017	0.95	1.10
60	0.362	0.016	0.89	1.42
61	0.894	0.017	1.00	0.95
62	0.631	0.016	0.94	1.16
63	1.315	0.018	1.09	0.74
64	1.785	0.019	1.06	0.90
65	1.534	0.018	1.00	0.90
66	0.276	0.016	0.97	1.08
67	1.432	0.018	0.90	1.13
68	0.595	0.016	0.91	1.27
69	0.734	0.016	1.08	0.70
70	1.346	0.018	0.97	1.01
71	2.090	0.021	1.04	0.78
72	1.989	0.020	0.75	1.52
73	2.704	0.025	0.78	1.01
74	2.240	0.022	0.99	0.82
75	2.404	0.023	1.31	0.42
76	1.802	0.019	1.09	0.72
77	1.968	0.020	1.29	0.35
78	2.525	0.024	1.15	0.65
79	2.240	0.022	1.10	0.73
80	2.339	0.022	1.30	0.55

Table 3

Characteristics of the anchor tests

Anchor Test No	Specifications	Actual Test Height: Without Person Fit	Actual Test Height: With t = 2.5 Person Fit
01.	All Anchor items (50L)	0.003	0.006
02.	Five best-fit items	0.272	0.288
03.	Ten best-fit items	0.062	0.072
04.	- 1.0 H / 1.0 W / 5 L	-0.976	-1.065
05.	- 1.0 H / 1.0 W / 10 L	-1.078	-1.083
06.	- 1.0 H / 2.0 W / 5 L	-0.999	-1.135
07.	- 1.0 H / 2.0 W / 10 L	-1.001	-1.108
08.	- 1.0 H / 2.0 W / 20 L	-0.999	-1.110
09.	0.0 H / 1.0 W / 5 L	0.017	0.014
10.	0.0 H / 1.0 W / 10 L	0.032	0.029
11.	0.0 H / 1.0 W / 20 L	0.023	0.030
12.	0.0 H / 2.0 W / 5 L	0.020	0.021
13.	0.0 H / 2.0 W / 10 L	0.012	0.014
14.	0.0 H / 2.0 W / 20 L	0.003	0.008
15.	0.0 H / 3.0 W / 5 L	0.036	0.068
16.	0.0 H / 3.0 W / 10 L	0.038	0.061
17.	0.0 H / 3.0 W / 20 L	0.033	-0.031
18.	1.0 H / 1.0 W / 5 L	1.022	1.142
19.	1.0 H / 1.0 W / 10 L	1.017	1.131
20.	1.0 H / 2.0 W / 5 L	0.983	1.086
21.	1.0 H / 2.0 W / 10 L	0.988	1.104
22.	1.0 H / 2.0 W / 20 L	0.991	1.104

Table 4

Items in Anchor Tests and Nonanchor Tests

Anchor Tests

01	2, 4, 5, 8, 9, 12-17, 19-21, 24-26, 28-30, 34, 35, 37, 37, 40, 42, 46, 49-52, 55-61, 63-71, 76, 77, 79
02	26, 40, 51, 55, 59
03	4, 26, 28, 30, 35, 40, 51, 55, 59, 67
04	9, 15, 17, 20, 42
05	8, 9, 14, 15, 16, 17, 19, 20, 21, 25
06	4, 12, 20, 28, 42
07	4, 5, 8, 12, 16, 19, 20, 21, 28, 49
08	2, 4, 5, 8, 9, 12, 13, 14, 15, 16, 17, 19, 21, 25, 26, 28, 30, 37, 42. 52
09	20, 34, 46, 49, 66
10	20, 24, 26, 37, 39, 40, 46, 51, 52, 60
11	20, 24, 26, 28, 29, 30, 34, 35, 37, 39, 40, 46, 49, 50, 51, 52, 55, 58, 60, 66
12	20, 34, 42, 50, 57
13	14, 17, 35, 37, 40, 42, 57, 60, 61, 68
14	8, 14, 17, 21, 26, 30, 37, 39, 40, 42, 50, 52, 55, 56, 57, 58, 60, 61, 68, 69
15	9, 17, 34, 61, 76
16	9, 17, 19, 20, 29, 50, 51, 61, 63, 76
17	8, 9, 14, 15, 19, 21, 25, 26, 30, 37, 40, 46, 58, 59, 61, 65, 66, 69, 70, 76
18	46, 57, 61, 70, 76
19	46, 56, 57, 59, 61, 63, 65, 67, 69, 76
20	34, 46, 57, 65, 79
21	34, 50, 55, 57, 61, 67, 68, 71, 76, 79
22	34, 40, 46, 55, 56, 57, 59, 60, 61, 63, 64, 65, 67, 68, 69, 70, 71, 76, 77, 79

Nonanchor Tests

Grade 6

3, 7, 22, 38, 43, 44, 45, 54, 62, 75, 1, 6, 10, 11, 18, 23, 27, 31, 33, 36

Grade 7

3, 7, 22, 38, 43, 44, 45, 54, 62, 75, 32, 41, 47, 48, 53, 72, 73, 74, 75, 80

TABLE 5
EVALUATION OF FIDELITY DISTANCE OVER 20 NONANCHOR ITEMS
GRADE 6 WITHOUT PERSON FIT

SMPL=1

TEST	HEIGHT	WIDTH	LENGTH	MEAN	SD	MIN	MAX
1				0.0786	0.0507	0.0001	0.1749
2				0.0776	0.0615	0.0096	0.1964
3				0.0757	0.0550	0.0106	0.1734
4	-1	1	5	0.0906	0.0708	0.0035	0.2475
5	-1	1	10	0.0978	0.0830	0.0022	0.3012
6	-1	2	5	0.0768	0.0710	0.0010	0.2360
7	-1	2	10	0.0784	0.0616	0.0093	0.2157
8	-1	2	20	0.0846	0.0703	0.0079	0.2339
9	0	1	5	0.0844	0.0686	0.0010	0.2460
10	0	1	10	0.0764	0.0596	0.0094	0.2044
11	0	1	20	0.0781	0.0493	0.0103	0.1657
12	0	2	5	0.1183	0.0814	0.0096	0.2984
13	0	2	10	0.0825	0.0682	0.0049	0.2449
14	0	2	20	0.0793	0.0556	0.0046	0.2016
15	0	3	5	0.0756	0.0605	0.0004	0.1854
16	0	3	10	0.0792	0.0506	0.0118	0.1592
17	0	3	20	0.0880	0.0661	0.0043	0.2227
18	1	1	5	0.1246	0.0856	0.0002	0.3062
19	1	1	10	0.0797	0.0525	0.0048	0.1868
20	1	2	5	0.1239	0.0853	0.0012	0.3068
21	1	2	10	0.1036	0.0687	0.0238	0.2438
22	1	2	20	0.0804	0.0513	0.0010	0.1700

SMPL=2

TEST	HEIGHT	WIDTH	LENGTH	MEAN	SD	MIN	MAX
1				0.0720	0.0536	0.0001	0.1589
2				0.0840	0.0530	0.0092	0.1642
3				0.0692	0.0513	0.0012	0.1378
4	-1	1	5	0.0834	0.0555	0.0001	0.1851
5	-1	1	10	0.1363	0.0756	0.0106	0.2726
6	-1	2	5	0.1027	0.0612	0.0136	0.2176
7	-1	2	10	0.0789	0.0532	0.0016	0.1656
8	-1	2	20	0.0745	0.0516	0.0063	0.1872
9	[illegible]	1	5	0.0930	0.0546	0.0052	0.1852
10	0	1	10	0.0826	0.0530	0.0080	0.1630
11	0	1	20	0.0698	0.0524	0.0013	0.1483
12	0	2	5	0.1083	0.0588	0.0238	0.2028
13	0	2	10	0.0801	0.0521	0.0064	0.1674
14	0	2	20	0.0919	0.0537	0.0067	0.1787
15	0	3	5	0.0901	0.0549	0.0034	0.1716
16	0	3	10	0.0745	0.0556	0.0009	0.1739
17	0	3	20	0.0695	0.0451	0.0161	0.1699
18	1	1	5	0.1102	0.0617	0.0250	0.2190
19	1	1	10	0.0856	0.0596	0.0083	0.2077
20	1	2	5	0.0707	0.0464	0.0120	0.1740
21	1	2	10	0.0734	0.0532	0.0079	0.1759
22	1	2	20	0.0717	0.0520	0.0020	0.1740

TABLE 5
EVALUATION OF FIDELITY DISTANCE OVER 20 NONANCHOR ITEMS
GRADE 6 WITHOUT PERSON FIT

------------------------------ SMPL=3 ------------------------------------

TEST	HEIGHT	WIDTH	LENGTH	MEAN	SD	MIN	MAX
1				0.0919	0.0585	0.0026	0.1756
2				0.1132	0.0649	0.0316	0.2556
3				0.0952	0.0606	0.0002	0.2072
4	-1	1	5	0.0948	0.0683	0.0023	0.2373
5	-1	1	10	0.1065	0.0822	0.0152	0.[illegible]
6	-1	2	5	0.0993	0.0712	0.0138	0.2[illegible]
7	-1	2	10	0.1027	0.0658	0.0030	0.2370
8	-1	2	20	0.0959	0.0726	0.0046	0.2466
9	0	1	5	0.0932	0.0595	0.0172	0.2078
10	0	1	10	0.0925	0.0595	0.0022	0.2088
11	0	1	20	0.0961	0.0593	0.0001	0.1980
12	0	2	5	0.0933	0.0582	0.0178	0.1888
13	0	2	10	0.0924	0.0605	0.0034	0.2116
14	0	2	20	0.0920	0.0576	0.0040	0.1759
15	0	3	5	0.0934	0.0596	0.0020	0.1880
16	0	3	10	0.1052	0.0555	0.0150	0.1960
17	0	3	20	0.1139	0.0613	0.0342	0.2442
18	1	1	5	0.0921	0.0640	0.0136	0.2174
19	1	1	10	0.0917	0.0614	0.0195	0.2205
20	1	2	5	0.0961	0.0855	0.0016	0.2434
21	1	2	10	0.1049	0.0578	0.0024	0.2056
22	1	2	20	0.0977	0.0589	0.0073	0.1987

TABLE 6
EVALUATION OF FIDELITY DISTANCE OVER 20 NONANCHOR ITEMS
GRADE 6 WITH PERSON FIT

SMPL=1

TEST	HEIGHT	WIDTH	LENGTH	MEAN	SD	MIN	MAX
1				0.0879	0.0671	0.0069	0.2609
2				0.1199	0.0815	0.0140	0.3820
3				0.1125	0.0818	0.0066	0.3944
4	-1	1	5	0.1197	0.0868	0.0002	0.3762
5	-1	1	10	0.1123	0.0765	0.0299	0.3411
6	-1	2	5	0.1030	0.0887	0.0034	0.2666
7	-1	2	10	0.1047	0.0877	0.0007	0.3237
8	-1	2	20	0.1081	0.0766	0.0155	0.3055
9	0	1	5	0.1040	0.0767	0.0152	0.3012
10	0	1	10	0.1074	0.0785	0.0026	0.3594
11	0	1	20	0.1197	0.0940	0.0027	0.4417
12	0	2	5	0.1208	0.0916	0.0198	0.2938
13	0	2	10	0.0949	0.0783	0.0015	0.2905
14	0	2	20	0.1037	0.0737	0.0253	0.3313
15	0	3	5	0.1061	0.0654	0.0294	0.2756
16	0	3	10	0.1095	0.0758	0.0146	0.3694
17	0	3	20	0.0973	0.0716	0.0086	0.2787
18	1	1	5	0.1270	0.0956	0.0196	0.3416
19	1	1	10	0.0986	0.0714	0.0128	0.3052
20	1	2	5	0.1534	0.1077	0.0076	0.3916
21	1	2	10	0.1229	0.0803	0.0299	0.2911
22	1	2	20	0.1022	0.0885	0.0016	0.2886

SMPL=2

TEST	HEIGHT	WIDTH	LENGTH	MEAN	SD	MIN	MAX
1				0.0819	0.0556	0.0026	0.2006
2				0.0850	0.0420	0.0210	0.1560
3				0.0822	0.0610	0.0019	0.2039
4	-1	1	5	0.0913	0.0536	0.0136	0.2044
5	-1	1	10	0.0884	0.0543	0.0048	0.1858
6	-1	2	5	0.1180	0.0849	0.0094	0.2844
7	-1	2	10	0.0889	0.0525	0.0066	0.1844
8	-1	2	20	0.0845	0.0422	0.0113	0.1513
9	0	1	5	0.0869	0.0435	0.0230	0.1520
10	0	1	10	0.0863	0.0491	0.0087	0.1833
11	0	1	20	0.0946	0.0711	0.0071	0.2781
12	0	2	5	0.0967	0.0554	0.0146	0.1794
13	0	2	10	0.0816	0.0423	0.0045	0.1675
14	0	2	20	0.0891	0.0551	0.0266	0.1964
15	0	3	5	0.0880	0.0489	0.0184	0.1704
16	0	3	10	0.0845	0.0552	0.0014	0.1724
17	0	3	20	0.0854	0.0507	0.0049	0.1768
18	1	1	5	0.1000	0.0633	0.0094	0.2014
19	1	1	10	0.0921	0.0630	0.0020	0.1860
20	1	2	5	0.0849	0.0515	0.0032	0.1852
21	1	2	10	0.0839	0.0543	0.0050	0.1620
22	1	2	20	0.0818	0.0555	0.0029	0.1740

TABLE 6
EVALUATION OF FIDELITY DISTANCE OVER 20 NONANCHOR ITEMS
GRADE 6 WITH PERSON FIT

SMPL=3

TEST	HEIGHT	WIDTH	LENGTH	MEAN	SD	MIN	MAX
1				0.1102	0.0595	0.0180	0.2250
2				0.1460	0.0919	0.0030	0.3540
3				0.1326	0.0728	0.0264	0.3036
4	-1	1	5	0.1133	0.0728	0.0026	0.2556
5	-1	1	10	0.1141	0.0717	0.0015	0.2485
6	-1	2	5	0.1131	0.0893	0.0024	0.3064
7	-1	2	10	0.1168	0.0690	0.0092	0.2528
8	-1	2	20	0.1097	0.0707	0.0028	0.2168
9	0	1	5	0.1109	0.0718	0.0030	0.2560
10	0	1	10	0.1264	0.0731	0.0105	0.2985
11	0	1	20	0.1498	0.0830	0.0236	0.3664
12	0	2	5	0.1156	0.0766	0.0036	0.2546
13	0	2	10	0.1106	0.0753	0.0095	0.2345
14	0	2	20	0.1267	0.0759	0.0353	0.3193
15	0	3	5	0.1089	0.0714	0.0008	0.2602
16	0	3	10	0.1313	0.0789	0.0071	0.3109
17	0	3	20	0.1136	0.0705	0.0070	0.2590
18	1	1	5	0.1122	0.0731	0.0118	0.2628
19	1	1	10	0.1153	0.0774	0.0007	0.2463
20	1	2	5	0.1329	0.1121	0.0018	0.3362
21	1	2	10	0.1217	0.0703	0.0085	0.2435
22	1	2	20	0.1126	0.0728	0.0056	0.2384

TABLE 7

EVALUATION OF FIDELITY DISTANCE OVER 20 NONANCHOR ITEMS

GRADE 7 WITHOUT PERSON FIT

SMPL=1

TEST	HEIGHT	WIDTH	LENGTH	MEAN	SD	MIN	MAX
1				0.1010	0.0717	0.0114	0.3226
2				0.1037	0.0756	0.0022	0.2848
3				0.0994	0.0728	0.0024	0.3236
4	-1	1	5	0.1079	0.0775	0.0085	0.2785
5	-1	1	10	0.1400	0.0929	0.0136	0.2864
6	-1	2	5	0.1174	0.0930	0.0004	0.3966
7	-1	2	10	0.1024	0.0763	0.0015	0.3395
8	-1	2	20	0.1045	0.0779	0.0173	0.3443
9	0	1	5	0.1017	0.0749	0.0110	0.3180
10	0	1	10	0.1016	0.0738	0.0050	0.3080
11	0	1	20	0.1017	0.0702	0.0081	0.3091
12	0	2	5	0.102[illegible]	0.0738	0.0040	0.3330
13	0	2	10	0.1021	0.0760	0.0172	0.3372
14	0	2	20	0.1000	0.0732	0.0032	0.3238
15	0	3	5	0.1259	0.0915	0.0028	0.2722
16	0	3	10	0.1078	0.0798	0.0036	0.2606
17	0	3	20	0.1131	0.0893	0.0109	0.3861
18	1	1	5	0.1041	0.0787	0.0084	0.3524
19	1	1	10	0.1093	0.0875	0.0080	0.3750
20	1	2	5	0.1045	0.0788	0.0080	0.3540
21	1	2	10	0.1018	0.0738	0.0079	0.3331
22	1	2	20	0.1039	0.0807	0.0063	0.3557

SMPL=2

TEST	HEIGHT	WIDTH	LENGTH	MEAN	SD	MIN	MAX
1				0.0995	0.0667	0.0120	0.2290
2				0.0988	0.0677	0.0104	0.2294
3				0.1023	0.0657	0.0064	0.2504
4	-1	1	5	0.1025	0.0716	0.0029	0.2741
5	-1	1	10	0.0991	0.0838	0.0024	0.2746
6	-1	2	5	0.1514	0.0977	0.0312	0.3808
7	-1	2	10	0.1149	0.0801	0.0137	0.3123
8	-1	2	20	0.1113	0.0795	0.0120	0.3130
9	0	1	5	0.1063	0.0713	0.0100	0.2840
10	0	1	10	0.0987	0.0681	0.0071	0.2301
11	0	1	20	0.0998	0.0644	0.0030	0.2240
12	0	2	5	0.1040	0.0671	0.0012	0.2612
13	0	2	10	0.0992	0.0768	0.0087	0.2563
14	0	2	20	0.0993	0.0678	0.0032	0.2398
15	0	3	5	0.1062	0.0893	0.0022	0.2872
16	0	3	10	0.1011	0.0733	0.0096	0.2464
17	0	3	20	0.1150	0.0790	0.0188	0.3062
18	1	1	5	0.0997	0.0806	0.0034	0.2664
19	1	1	10	0.0997	0.0688	0.0207	0.2383
20	1	2	5	0.0993	0.0709	0.0206	0.2426
21	1	2	10	0.1007	0.0781	0.0001	0.2599
22	1	2	20	0.0999	0.0744	0.0051	0.2519

TABLE 7
EVALUATION OF FIDELITY DISTANCE OVER 20 NONANCHOR ITEMS
GRADE 7 WITHOUT PERSON FIT

SMPL=3

TEST	HEIGHT	WIDTH	LENGTH	MEAN	SD	MIN	MAX
1				0.0982	0.0615	0.0223	0.2263
2				0.0974	0.0650	0.0072	0.2202
3				0.1039	0.0625	0.0269	0.2531
4	-1	1	5	0.1000	0.0714	0.0091	0.2379
5	-1	1	10	0.1050	0.0878	0.0024	0.2776
6	-1	2	5	0.1206	0.0749	0.0092	0.3102
7	-1	2	10	0.1142	0.0718	0.0035	0.2935
8	-1	2	20	0.1139	0.0705	0.0013	0.2917
9	0	1	5	0.1265	0.0762	0.0264	0.3166
10	0	1	10	0.0926	0.0698	0.0028	0.2248
11	0	1	20	0.0965	0.0622	0.0161	0.2131
12	0	2	5	0.1019	0.0614	0.0090	0.2460
13	0	2	10	0.1086	0.0645	0.0211	0.2709
14	0	2	20	0.0996	0.0616	0.0223	0.2313
15	0	3	5	0.0994	0.0612	0.0224	0.2336
16	0	3	10	0.0904	0.0796	0.0002	0.2422
17	0	3	20	0.1331	0.0805	0.0028	0.3262
18	1	1	5	0.1075	0.0665	0.0032	0.2702
19	1	1	10	0.1009	0.0613	0.0132	0.2418
20	1	2	5	0.1210	0.0719	0.0212	0.3012
21	1	2	10	0.0999	0.0613	0.0056	0.2394
22	1	2	20	0.1003	0.0609	0.0110	0.2400

TABLE 8
EVALUATION OF FIDELITY DISTANCE OVER 20 NONANCHOR ITEMS
GRADE 7 WITH PERSON FIT

SMPL=1

TEST	HEIGHT	WIDTH	LENGTH	MEAN	SD	MIN	MAX
1				0.1177	0.0749	0.0365	0.3545
2				0.1242	0.0923	0.0200	0.3010
3				0.1157	0.0751	0.0214	0.3094
4	-1	1	5	0.1118	0.0886	0.0116	0.3094
	1	1	10	0.1155	0.0958	0.0056	0.3616
6	-1	2	5	0.1304	0.0986	0.0010	0.4360
7	-1	2	10	0.1170	0.0722	0.0312	0.3468
8	-1	2	20	0.1164	0.0766	0.0288	0.3308
9	0		5	0.1201	0.0815	0.0214	0.3374
10	0		10	0.1225	0.0848	0.0065	0.3075
11	0	1	20	0.1242	0.0848	0.0049	0.3059
12	0	2	5	0.1149	0.0789	0.0268	0.3552
13	0	2	10	0.1091	0.0809	0.0075	0.3565
14	0	2	20	0.1147	0.0771	0.0228	0.3248
15	0	3	5	0.1384	0.1158	0.0018	0.3508
16	0	3	10	0.1284	0.0984	0.0112	0.3098
17	0	3	20	0.1175	0.0824	0.0221	0.3711
18	1	1	5	0.1259	0.0873	0.0040	0.3760
19	1	1	10	0.1361	0.1017	0.0114	0.4204
20	1	2	5	0.1226	0.0873	0.0008	0.3948
21	1	2	10	0.1173	0.0884	0.0082	0.3838
22	1	2	20	0.1215	0.0919	0.0090	0.405[illegible]

SMPL=2

TEST	HEIGHT	WIDTH	LENGTH	MEAN	SD	MIN	MAX
1				0.1094	0.0694	0.0098	0.2218
2				0.1046	0.0662	0.0084	0.2126
3				0.1044	0.0679	0.0160	0.2380
4	-1	1	5	0.1088	0.0697	0.0062	0.2728
5	-1	1	10	0.1170	0.0815	0.0187	0.3377
6	-1	2	5	0.1315	0.0885	0.0048	0.3302
7	-1	2	10	0.1152	0.0607	0.0377	0.2537
8	-1	2	20	0.1070	0.0662	0.0022	0.2478
9	0	1	5	0.1102	0.0647	0.0172	0.2602
10	0	1	10	0.1075	0.0748	0.0091	0.2741
11	0	1	20	0.1133	0.0802	0.0015	0.2775
12	0	2	5	0.1025	0.0677	0.0024	0.2356
13	0	2	10	0.1038	0.0781	0.0017	0.2287
14	0	2	20	0.1016	0.0702	0.0024	0.2034
15	0	3	5	0.1070	0.0878	0.0110	0.2630
16	0	3	10	0.1124	0.0777	0.0013	0.2463
17	0	3	20	0.1082	0.0681	0.0030	0.2270
18	1	1	5	0.1037	0.0705	0.0038	0.2328
19	1	1	10	0.1090	0.0749	0.0025	0.2625
20	1	2	5	0.1027	0.0642	0.0040	0.2220
21	1	2	10	0.1110	0.0684	0.0068	0.2312
22	1	2	20	0.1124	0.0676	0.0060	0.2300

TABLE 8

EVALUATION OF FIDELITY DISTANCE OVER 20 NONANCHOR ITEMS

GRADE 7 WITH PERSON FIT

SMPL=3

TEST	HEIGHT	WIDTH	LENGTH	MEAN	SD	MIN	MAX
1				0.0892	0.0638	0.0046	0.1996
2				0.0978	0.0685	0.0090	0.2470
3				0.0952	0.0660	0.0011	0.2531
4	-1	1	5	0.0991	0.0711	0.0092	0.2542
5	-1	1	10	0.1128	0.0719	0.0061	0.3069
6	-1	2	5	0.1169	0.0707	0.0172	0.3262
7	-1	2	10	0.1103	0.0667	0.0008	0.2822
8	-1	2	20	0.1002	0.0704	0.0069	0.2671
9	0	1	5	0.1262	0.0812	0.0168	0.3768
10	0	1	10	0.0957	0.0734	0.0155	0.2575
11	0	1	20	0.0913	0.0676	0.0013	0.2227
12	0	2	5	0.0997	0.0698	0.0008	0.2688
13	0	2	10	0.1052	0.0632	0.0063	0.2753
14	0	2	20	0.0901	0.0596	0.0035	0.1955
15	0	3	5	0.1031	0.0685	0.0020	0.2740
16	0	3	10	0.0919	0.0688	0.0142	0.2478
17	0	3	20	0.0962	0.0618	0.0129	0.2361
18	1	1	[illegible]	0.1262	0.0756	0.0190	0.3330
19	1	1	10	0.1054	0.0749	0.0026	0.2904
20	1	2	5	0.1287	0.0843	0.0162	0.3638
21	1	2	10	0.1035	0.0735	0.0040	0.2770
22	1	2	20	0.0978	0.0722	0.0060	0.2590