Navy Personnel Research and Development Center
San Diego, California 92152-7250 TN-94-4 October 1993

2

2

Reducing Test Length With Polychotomous Scoring

J. Bradford Sympson
Mark L. Davison

NPRDC-TN-94-4 October 1993

Reducing Test Length With Polychotomous Scoring

J. Bradford Sympson

Mark L. Davison
University of Minnesota
Minneapolis, Minnesota 55455-0296

Reviewed by
Daniel O. Segall

Approved and released by
W. A. Sands
Director, Personnel Systems Department

Navy Personnel Research and Development Center
San Diego, California 92152-7250

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE October 1993	3 REPORT TYPE AND DATE COVERED Final—October 1988-September 1991

4. TITLE AND SUBTITLE
Reducing Test Length With Polychotomous Scoring

5. FUNDING NUMBERS
Program Element: 0601153N
Work Unit: R4204

6. AUTHOR(S)
J. Bradford Sympson, Mark L. Davison

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Navy Personnel Research and Development Center
San Diego, California 92152-7250

8. PERFORMING ORGANIZATION REPORT NUMBER
NPRDC-94-4

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Office of the Assistant Secretary of Defense
The Pentagon
Washington, DC 20301-3210

10. SPONSORING/MONITORING

11. SUPPLEMENTARY NOTES
Functional Area: Personnel
Product Line: Computerized Testing
Effort: Computerized Adaptive Testing (CAT)

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE
A

13. ABSTRACT *(Maximum 200 words)*

This study compared two new types of test scoring (polyweighting and dichotomous Item Response Theory [IRT] scoring) t traditional number-correct (NC) scoring. The study used data collected from applicants for military enlistment who had taken th *Armed Services Vocational Aptitude Battery* (ASVAB). The objective was to determine the impact of these two new scoring methods on alternate-form reliabilities. Content areas studied were ASVAB Mathematics Knowledge (MK) and General Science (GS).

Examinees in one part of the study completed one of five experimental MK tests and one of six operational MK tests. Othe examinees in the study completed one of four experimental GS tests and one of six operational GS tests. Experimental and operational tests were analyzed separately. For each type of test, at each possible test length, median alternate-form reliability was determined for each scoring method.

The authors reached five conclusions: (1) Polyweighting is superior to NC scoring; (2) dichotomous IRT scoring is usually superior to NC scoring, but not always; (3) when dichotomous IRT scoring works well, polyweighting and dichotomous IRT scoring allow similar reductions in test length; (4) polyweighting and IRT scoring both provide greater benefits when item difficulties are more variable; and (5) polyweighting and IRT scoring use different (complementary) mechanisms for increasing measurement precision.

14. SUBJECT TERMS
Selection, classification, training, testing, item scoring, polychotomous scoring, polytomous scoring

15. NUMBER OF PAGES
29

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT UNCLASSIFIED	18. SECURITY CLASSIFICATION OF THIS PAGE UNCLASSIFIED	19. SECURITY CLASSIFICATION OF ABSTRACT UNCLASSIFIED	20. LIMITATION OF ABSTRACT UNLIMITED

Foreword

This technical note describes a study designed to evaluate the potential of two new multiple-choice item scoring methods for reducing test length. Both methods produced higher levels of alternate-form reliability than the more traditional, number-correct scoring method. This increased reliability can be translated into decreased test length required to achieve a target level of reliability.

Results reported in this technical note were originally presented at the annual meeting of the American Educational Research Association, held in San Francisco, March 1989. It is being published at this time for archival purposes.

The research described here was funded by the Office of Naval Research (Code 1142) under the Navy Laboratory Participation Program (Program Element 0601153N, Work Order R4204). Additional funding was provided by the Joint Service Computerized Adaptive Testing-Armed Services Vocational Aptitude Battery (CAT-ASVAB) Program, which is sponsored by the Office of the Assistant Secretary of Defense (FM&P).

W. A. SANDS
Director, Personnel Systems Department

Summary

Problem

Conventional methods for scoring aptitude and achievement tests that are used in selecting, classifying, and training military personnel discard useful information about an examinee's ability/skill level. Information is lost whenever the original responses to test questions are classified only as "right" or "wrong." Additional information can be obtained by considering the difficulty level of the questions answered correctly and by taking into account which particular wrong answers were selected.

Objective

The objective of this effort was to develop new procedures for scoring aptitude and achievement tests that will increase the reliability and validity of those tests.

Approach

This study compared two new types of test scoring (polyweighting and dichotomous Item Response Theory [IRT] scoring) to traditional number-correct (NC) scoring. The study used data collected from applicants for military enlistment who had taken the Armed Services Vocational Aptitude Battery (ASVAB). The objective was to determine the impact of these two new scoring methods on alternate-form reliabilities. Content areas studied were ASVAB Mathematics Knowledge (MK) and General Science (GS).

Results

The authors reached five conclusions: (1) Polyweighting is superior to NC scoring; (2) dichotomous IRT scoring is usually superior to NC scoring, but not always; (3) when dichotomous IRT scoring works well, polyweighting and dichotomous IRT scoring allow similar reductions in test length; (4) polyweighting and IRT scoring both provide greater benefits when item difficulties are more variable; and (5) polyweighting and IRT scoring use different (complementary) mechanisms for increasing measurement precision.

Conclusion

Results of this study indicate that polyweighting and dichotomous IRT scoring both increase alternate-form reliability, which in turn provides an opportunity to reduce test length without sacrificing reliability. However, dichotomous IRT scoring did fail in one of the conditions studied. Polyweighting is recommended for use in applications where sample sizes range from $N = 100$ to $N = 1{,}000$ and/or the items to be calibrated are multidimensional.

Recommendation

Organizations that administer aptitude and/or achievement tests for purposes of personnel selection, classification, or training should consider whether polyweighting and/or IRT scoring can be usefully applied to their tests.

Contents

List of Figures

Introduction

In a review of research on polychotomous item scoring, Haladyna and Sympson (1988) distinguished between two broad classes of scoring procedures. One approach involves the assignment of differential scoring weights to item response categories. In this approach, the test score is a linear function of the examinee's item response vector. The second approach to polychotomous item scoring discussed by Haladyna and Sympson is based on item response theory (IRT). In this approach, the test score is a nonlinear function of the examinee's item response vector.

Sympson (1993b) has introduced a new type of linear polychotomous scoring referred to as *polyweighting*. Polyweighting has three characteristics that distinguish it from most other methods of linear polychotomous scoring: (1) Polyweighting utilizes scoring weights that are independent of the difficulty of the items that are calibrated, (2) the scoring weights are bounded so that no wrong answer to a question ever gets a higher weight than the correct answer, and (3) data-sets in which different examinees have been administered different questions can be used for item calibration. Computation of *polyweights* is accomplished using the computer program POLY.

Sympson and Haladyna (1993) conducted an empirical evaluation of polyweighting in the context of medical certification testing. In that study, data from 1,100 resident physicians who had completed a 200-item test in the field of otolaryngology (the diagnosis and treatment of ear, nose, and throat disorders) were obtained. Five-hundred of these physicians were selected at random to make up "Sample A." Five-hundred different physicians were selected at random to make up "Sample B." Summary statistics and polyweights were determined for all 200 items in Sample A using the program POLY.

Using the set of 200 items as an item bank, Sympson and Haladyna assembled 20 short (10-, 20-, 30-, 40-item) assessment tests and scored them in Sample B. Twelve assessment tests were assembled by randomly selecting items and eight assessment tests were assembled by selecting "best" items. Both proportion-correct scores and test scores based on the Sample A polyweights were computed in Sample B. Then, both types of test scores were correlated with Sample B 200-item domain scores.

For all 20 assessment tests, polyweighting resulted in higher test reliability (coefficient-α) and domain validity in Sample B. The observed increases in reliability corresponded to a mean increase in test length of 28%. Over all 20 tests, the mean increase in domain validity was .075. The minimum increase in domain validity was .052.

The results of the Sympson and Haladyna (1993) study suggest that polyweighting should allow reductions in test length, while maintaining test reliability at the level observed under traditional number/proportion-correct scoring. This is implied by the increases in coefficient-α that were observed. In the study reported in this paper, we determine how much tests can be shortened without reducing alternate-form reliability if they are scored using polyweighting rather than number-correct scoring. We also determine how much tests can be shortened if they are scored using a dichotomous IRT model.

A Brief Description of Polyweighting

The examinee scores obtained when using polyweights to score a test are called *polyscores*. An examinee's polyscore is equal to the mean of the polyweights of the categories chosen by the examinee.

An iterative procedure is used to derive polyweights for a set of items. The procedure is described in Sympson (1993b). Polyweights are defined as follows:

1. For each correct answer, the polyweight is equal to the mean percentile rank among examinees choosing the answer, rounded to the nearest integer.

2. For each wrong answer chosen by 100 or more examinees, the provisional polyweight is equal to the mean percentile rank among examinees choosing the answer, rounded to the nearest integer.

3. For each wrong answer chosen by fewer than 100 examinees, the provisional polyweight is a rounded linear combination of the mean percentile rank among examinees choosing the answer and the mean percentile rank among examinees choosing any wrong answer on the item. For these response categories, the polyweight for category *j* of item *i* is equal to

$$W_{ij} = \overline{R}_{i(w)} + \left[\frac{Nij}{100}\right]^{1/2} \overline{R}_{ij} - \overline{R}_{i(w)} \, , \tag{1}$$

rounded to the nearest integer. In Equation 1, $\overline{R}_{i(w)}$ is the mean percentile rank among examinees choosing any wrong answer on item *i*, R_{ij} is the mean-percentile rank among examinees choosing category *j*, and N_{ij} is the number of examinees choosing category *j*.

4. For a given item, if the provisional polyweight for an incorrect response is less than the polyweight for the correct response, the provisional polyweight is used as the category polyweight. However, if the provisional polyweight for an incorrect response equals or exceeds the polyweight for the correct response, the polyweight for the incorrect response is set equal to 1 less than the polyweight for the correct response.

In the program POLY, examinee percentile ranks range from a minimum possible value of 100(1/*N*) to a maximum possible value of 100 (where *N* is the number of examinees in the item calibration sample). Thus, polyweights can assume any integer value from 0 to 100.

Polyweighting is not based on IRT, and does not require any assumptions regarding "latent" abilities, the dimensionality of the set(s) of items analyzed, or the mathematical form of the regression of item responses on unobservable variables. The procedure does assume that the individuals included in an item analysis are randomly sampled from the examinee population of interest. The procedure is characterized as "linear" because each examinee's score is a linear function of category scoring weights and category-response indicators.

Unlike some scoring methods, polyweighting gives the examinee more credit for correct answers to difficult questions and less credit for correct answers to easy questions. Also, polyweighting penalizes the examinee more heavily for wrong answers to easy questions than for wrong answers to difficult questions. This may be contrasted with number/proportion-correct scoring and with scoring under the 1-parameter (Rasch) and 2-parameter logistic IRT models. The latter scoring methods assign scores to examinees in a manner that renders the scores independent of the difficulty of the questions answered correctly or incorrectly (Birnbaum, 1968, p. 458).

Method

In order to compare number-correct scoring, polyweighting, and IRT scoring in terms of their impact on alternate-form reliability, test data from applicants for military enlistment were used. These data were collected under a U.S. Air Force Armstrong Laboratory Human Resource Directorate (AL/HR) contract to prepare and calibrate test items for the Joint-Service Computerized Adaptive Testing-Armed Services Vocational Aptitude Battery (CAT-ASVAB) Project (Prestwood, Vale, Massey, & Welsh, 1985).

Two content areas that are included in the Armed Services Vocational Aptitude Battery (ASVAB) were studied: Mathematics Knowledge (MK) and General Science (GS). The MK item-bank included both 4- and 5-option multiple-choice items measuring general mathematics knowledge, including algebra and geometry. The GS item-bank included both 4- and 5-option multiple-choice items that measure knowledge of the physical, biological, and earth sciences. These two item-banks were selected because their content is similar to achievement tests used in educational settings.

Since one of the objectives of this study was to compare polyweighting with test scoring based on the 3-parameter logistic (3PL) IRT model (Birnbaum, 1968, p. 405), and since fitting the 3PL model requires a large sample of examinees (1,000 or more, if possible), this study was conducted with relatively large samples. Thus, the results reported here are indicative of the performance of the three scoring methods in high-volume testing programs.

Item Calibrations for Mathematics Knowledge

The MK database contained item responses from applicants for military enlistment who had been administered one of five experimental MK tests that contained 46 items each, and one of six operational MK tests that contained 25 items each. Thus, each examinee in the MK database had been administered 1 of 30 different item-sets, each containing 71 items.

Although six different operational MK tests were administered, these tests contained a total of only 50 unique items. Operational tests 1 and 2 contained the same 25 items, presented in a different order, and operational tests 3, 4, 5, and 6 all contained the same 25 items, presented in four different orders. Since there was evidence that the operational MK tests were slightly speeded, and since this could result in a given item having different polyweights and/or IRT parameters as a function of the item's position in a test, the MK calibrations were conducted as if a total of 380 unique items had been administered.

The MK examinees were initially assigned to 30 groups, based on the item-sets they had completed. After eliminating examinees who had responded to fewer than 40% of the items administered (i.e., 28 or fewer items), the remaining examinees in each group were allocated to a "calibration sample" and a "holdout sample" in a serial, odd-even fashion. Then, data from all examinees who had been assigned to a calibration sample were combined to create an "MK joint-calibration sample." The MK joint-calibration sample contained 6,447 examinees. The 30 holdout samples, taken together, contained 6,434 examinees.

The MK joint-calibration sample was used with the program POLY to compute polyweights for all MK items in a single POLY run. The same sample was used with the program LOGIST5 (Wingersky, Barton, & Lord, 1982) to compute 3PL item parameter estimates for these items in a single LOGIST run. The number of examinees in the MK joint-calibration sample that had been administered each MK test is given in Table 1. The exact number of examinees used for calibrating each MK item varied as a function of: (1) the test the item appeared in, (2) the calibration program used (LOGIST5 does not use examinees who don't reach an item), and (3) the position of the item in the test.

Table 1

Administration Frequencies for Mathematics Knowledge Tests

Test Type	Test ID	Frequency
Experimental	Test 1	1,353
	Test 2	1,322
	Test 3	1,305
	Test 4	1,251
	Test 5	1,216
Operational	Test 1	1,128
	Test 2	1,195
	Test 3	1,101
	Test 4	1,131
	Test 5	986
	Test 6	906

Mathematics Knowledge Alternate Forms

The data from the 30 MK holdout samples were used first to form five experimental-test holdout samples, one for each experimental MK test. Within each of these five samples, the 46-item experimental MK test was split into two 23-item forms by assigning items to alternate forms "A" and "B" in a sequential, "ABBAABB . . . A" pattern. This pattern was used so that one form would not contain items that were always administered before the items in the other form. Splitting the five experimental MK tests resulted in five pairs of alternate forms. Each examinee in an experimental test holdout sample had completed one of these pairs.

The data from the 30 MK holdout samples were also used to form six operational test holdout samples, one for each operational MK test. Within each of these six samples, the 25-item operational MK test was split into two 12-item forms by skipping the first item and assigning the remaining items to alternate forms "A" and "B" in a sequential, "ABBAABB . . . A" pattern. Splitting the six operational MK tests resulted in six pairs of alternate forms. Each examinee in an operational test holdout sample had completed one of these pairs.

Scoring Mathematics Knowledge Alternate Forms

For each of the MK alternate forms, three different test scores were computed for each holdout sample examinee at each possible test length (from 1 to 23 items for the experimental alternate-forms and from 1 to 12 items for the operational alternate forms). The three scores computed at each test length were: (1) number-correct (NC), (2) a polyscore computed using the polyweights obtained in the MK calibration sample, and (3) a Bayesian IRT ability estimate (Owen, 1975) computed using the 3PL parameters obtained in the MK calibration sample.

Assessing Test Length Reductions for Mathematics Knowledge

To assess potential reductions in test length that might be available by using polyweighting or IRT scoring, an alternate-form correlation was calculated for each type of test score at each possible test length for each pair of alternate forms. For each scoring method, this resulted in five alternate-form correlations at each test length from 1 to 23 items for the experimental MK alternate forms and six alternate-form correlations at each test length from 1 to 12 items for the operational MK alternate forms.

The first step in summarizing the results obtained for MK was to compute and plot the median alternate-form correlation for each scoring method at each possible test length. This was done separately for the experimental and operational alternate forms. Next, the median alternate-form correlation for NC scores at each possible test length was noted and the test length needed under polyweighting and under IRT scoring to obtain the same median alternate-form correlation was determined. To equate median alternate-form correlations under polyweighting and IRT scoring to each median correlation observed under NC scoring, linear interpolation between observed median correlations was used when computing the test lengths for each alternative scoring method (polyweighting or IRT).

To determine the proportionate test length reduction available under the alternative scoring method, the difference between the NC test length and the test length having equal median reliability under the alternative scoring method was divided by the NC test length. This index of proportionate reduction in test length was computed and plotted for polyscores and for IRT scores at each possible test length. This was done separately for experimental and operational alternate-forms.

Item Calibrations for General Science (GS)

The GS database contained item responses from applicants for military enlistment who had been administered one of four experimental GS tests that contained 57 items each, and one of six operational GS tests that contained 25 items each. Thus, each examinee in the GS database had been administered 1 of 24 different item-sets, each containing 82 items.

Although six different operational GS tests were administered, these tests contained a total of only 50 unique items. Operational tests 1 and 2 contained the same 25 items, presented in a different order, and operational tests 3, 4, 5, and 6 all contained the same 25 items, presented in four different orders. Since there was evidence that the operational GS tests were slightly speeded, and since this could result in a given item having different polyweights and/or IRT parameters as a function of the item's position in a test, the GS calibrations were conducted as if a total of 378 unique items had been administered.

The GS examinees were initially assigned to 24 groups, based on the item-sets they had completed. After eliminating examinees who had responded to fewer than 40% of the items administered (i.e., 32 or fewer items), the remaining examinees in each group were allocated to a calibration sample and a holdout sample in a serial, odd-even fashion. Then, data from all examinees who had been assigned to a calibration sample were combined to create a "GS joint-calibration sample." The GS joint-calibration sample contained 5,412 examinees. The 24 holdout samples, taken together, contained 5,398 examinees.

The GS joint-calibration sample was used with the program POLY to compute polyweights for all GS items in a single POLY run. The same sample was used with the program LOGIST5 to compute 3PL item parameter estimates for these items in a single LOGIST run. The number of examinees in the GS joint-calibration sample that had been administered each GS test is given in Table 2. The exact number of examinees used for calibrating each GS item varied as a function of: (1) the test the item appeared in, (2) the calibration program used, and (3) the position of the item in the test.

Table 2

Administration Frequencies for General Science Tests

Test Type	Test ID	Frequency
Experimental	Test 1	1,450
	Test 2	1,361
	Test 3	1,315
	Test 4	1,286
Operational	Test 1	1,005
	Test 2	1,005
	Test 3	917
	Test 4	937
	Test 5	780
	Test 6	768

General Science Alternate Forms

The data from the 24 GS holdout samples were used first to form four experimental-test holdout samples, one for each experimental GS test. Within each of these four samples, the 57-item experimental GS test was split into two 28-item forms by skipping the first item and assigning the remaining items to alternate forms "A" and "B" in a sequential, "ABBAABB . . . A" pattern.

Splitting the four experimental GS tests resulted in four pairs of alternate forms. Each examinee in an experimental test holdout sample had completed one of these pairs.

The data from the 24 GS holdout samples were also used to form six operational test holdout samples, one for each operational GS test. Within each of these six samples, the 25-item operational GS test was split into two 12-item forms by skipping the first item and assigning the remaining items to alternate forms "A" and "B" in a sequential, "ABBAABB . . . A" pattern. Splitting the six operational GS tests resulted in six pairs of alternate forms. Each examinee in an operational test holdout sample had completed one of these pairs.

Scoring General Science Alternate Forms

For each of the GS alternate forms created, three different test scores were computed for each holdout sample examinee at each possible test length (from 1 to 28 items for the experimental alternate forms and from 1 to 12 items for the operational alternate forms). The three scores computed at each test length were: (1) number correct, (2) a polyscore computed using the polyweights obtained in the GS calibration sample, and (3) a Bayesian IRT ability estimate (Owen, 1975) computed using the 3PL parameters obtained in the GS calibration sample.

Assessing Test Length Reductions for General Science

To assess potential reductions in test length that might be available by using polyweighting or IRT scoring, an alternate-form correlation was calculated for each type of test score at each possible test length for each pair of alternate forms. For each scoring method, this resulted in four alternate-form correlations at each test length from 1 to 28 items for the experimental GS alternate forms and six alternate-form correlations at each test length from 1 to 12 items for the operational GS alternate forms.

The first step in summarizing the results obtained for GS was to compute and plot the median alternate-form correlation for each scoring method at each possible test length. This was done separately for the experimental and operational alternate forms. Next, the median alternate-form correlation for NC scores at each possible test length was noted and the test length needed under polyweighting and under IRT scoring to obtain the same median alternate-form correlation was determined. To equate median alternate-form correlations under polyweighting and IRT scoring to each median correlation observed under NC scoring, linear interpolation between observed median correlations was used when computing the test lengths for each alternative scoring method (polyweighting or IRT).

To determine the proportionate test length reduction available under the alternative scoring method, the difference between the NC test length and the test length having equal median reliability under the alternative scoring method was divided by the NC test length. This index of proportionate reduction in test length was computed and plotted for polyscores and for IRT scores at each possible test length. This was done separately for experimental and operational alternate forms.

Results

Mathematics Knowledge Experimental Tests

Figure 1 is a plot of the median alternate-form correlation for polyscores and for NC scores as a function of test length in the experimental MK alternate forms. Figure 2 compares the median alternate-form correlations obtained for IRT scores with those obtained for NC scores. Figure 3 compares polyscores and IRT scores.

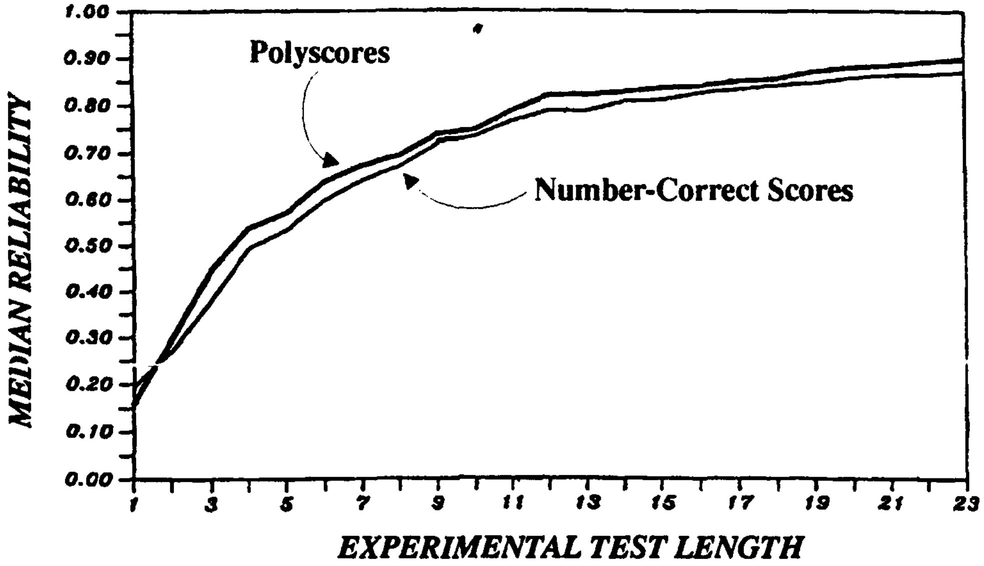

Figure 1. Median reliability for polyscores and number-correct scores as a function of experimental test length, for Mathematics Knowledge.

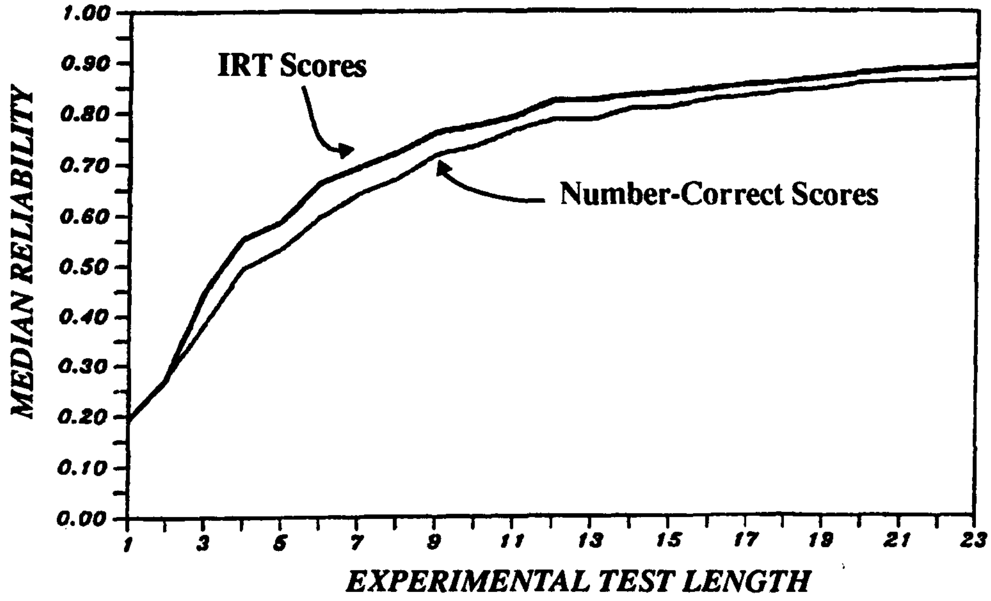

Figure 2. Median reliability for IRT scores and number-correct scores as a function of experimental test length, for Mathematics Knowledge.

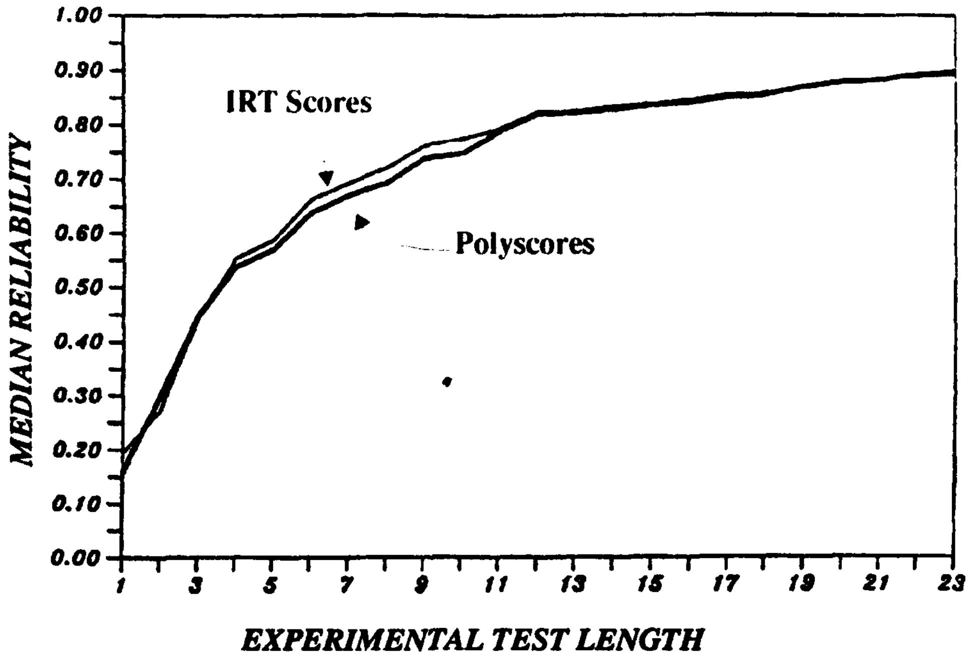

Figure 3. Median reliability for IRT scores and polyscores as a function of experimental test length, for Mathematics Knowledge.

Figure 4 shows the proportionate reduction in test length that could be realized, without reducing alternate-form reliability below the median value observed with NC scoring, under polyweighting and under IRT scoring, as a function of test length in the experimental MK alternate forms.

The following example illustrates how the values plotted in Figure 4 were obtained. At a test length of 15 items, the median alternate-form correlation for NC scores from the experimental MK alternate forms is .811. The median alternate-form correlation between polyscores computed on these same tests is .787 at a test length of 11 items and .819 at a test length of 12 items. Linear interpolation between the median correlations obtained for polyscores suggests that a test length of 11.73 items would provide an alternate-form correlation of .811. Dividing 15 minus 11.73 by 15 gives the indicated proportionate reduction in test length of .218 that is plotted for polyscores at a test length of 15 items. This result suggests that polyweighting offers the possibility of a 22% reduction in test length for 15-item experimental MK tests. Of course, this value is subject to sampling error and should be interpreted in conjunction with the other values plotted in Figure 4.

Mathematics Knowledge Operational Tests

Figure 5 is a plot of the median alternate-form correlation for polyscores and for NC scores as a function of test length in the operational MK alternate forms. Figure 6 compares the median alternate-form correlations obtained for IRT scores with those obtained for NC scores. Figure 7 compares polyscores and IRT scores.

Figure 8 shows the proportionate reduction in test length that could be realized, without reducing alternate-form reliability below the median value observed with NC scoring, under polyweighting and under IRT scoring, as a function of test length in the operational MK alternate forms.

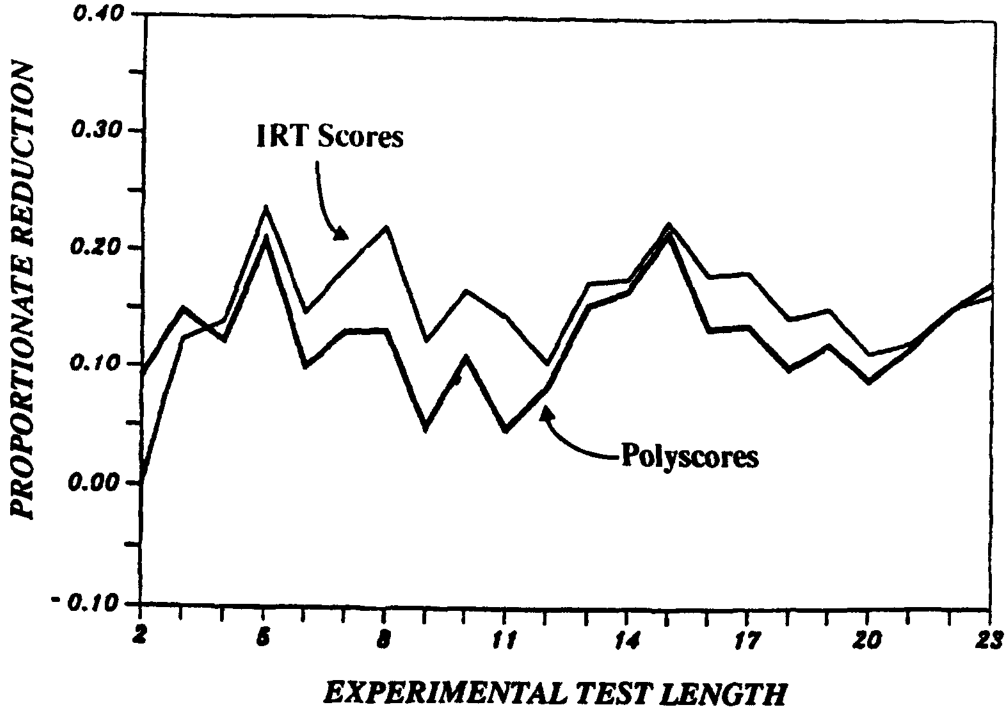

Figure 4. Proportionate reduction in test length under polyweighting and IRT scoring for the experimental Mathematics Knowledge forms.

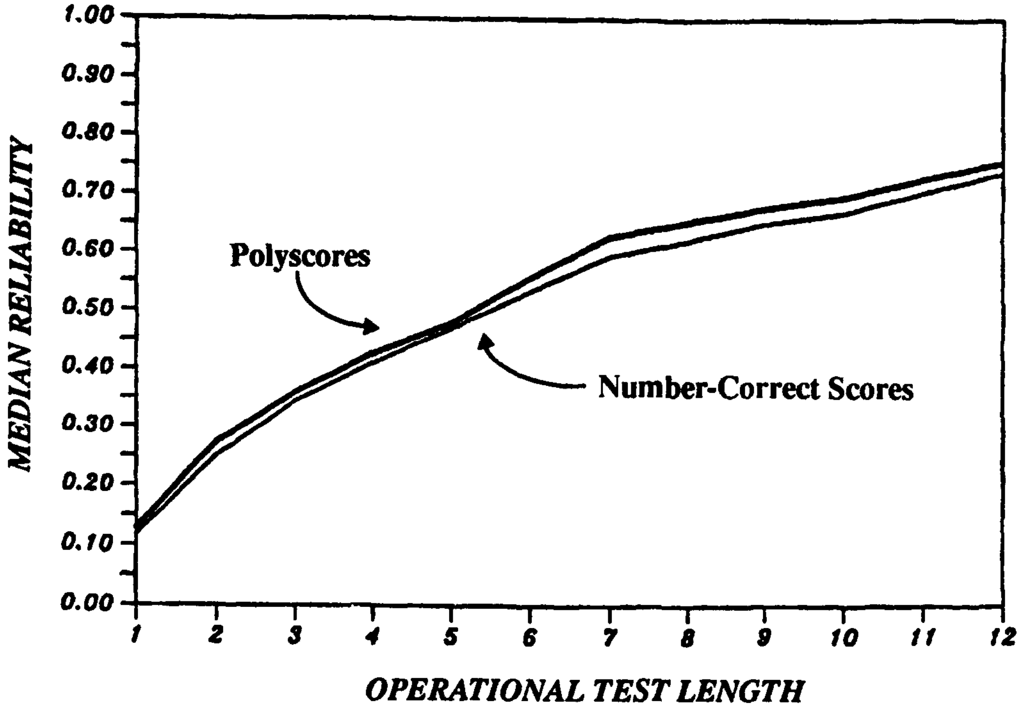

Figure 5. Median reliability for polyscores and number-correct scores as a function of operational test length in the Mathematics Knowledge alternate forms.

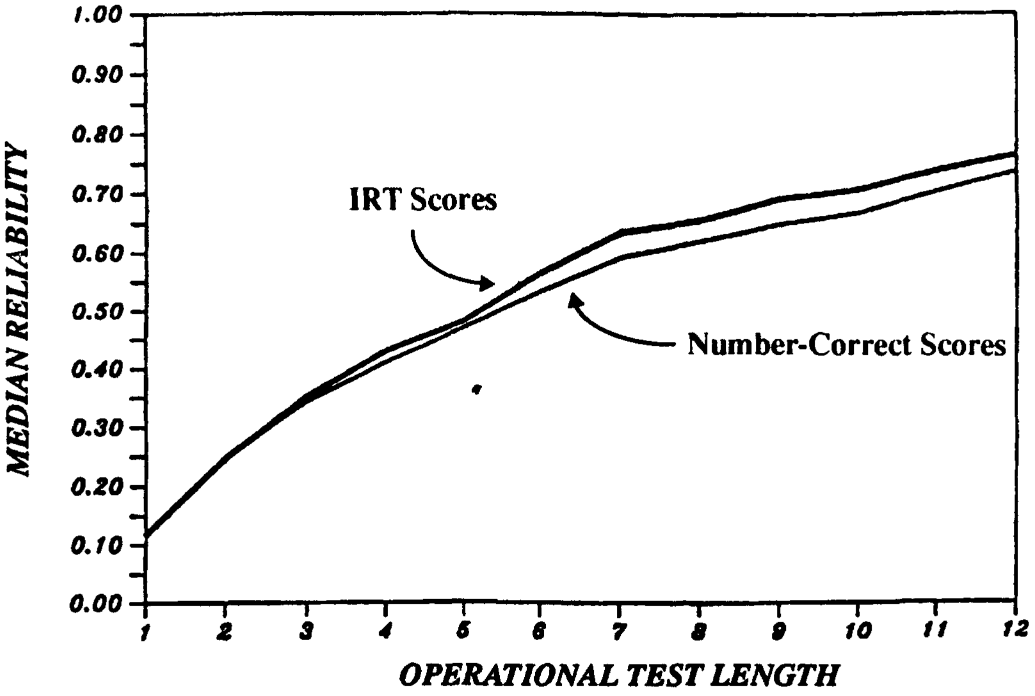

Figure 6. Median reliability for IRT scores and number-correct scores as a function of operational test length for Mathematics Knowledge.

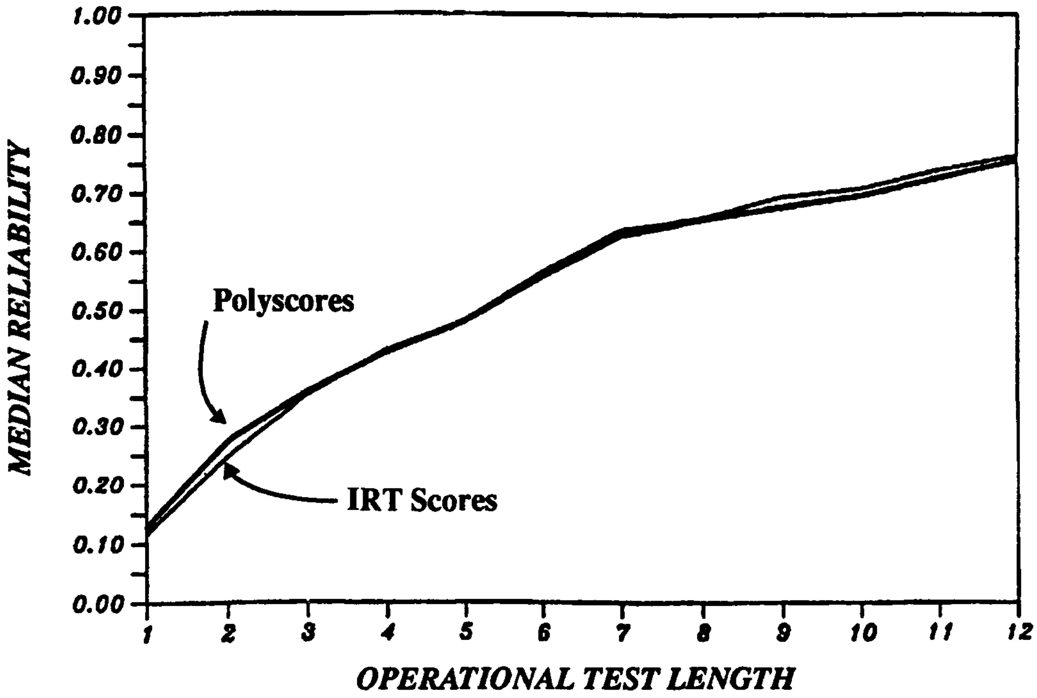

Figure 7. Median reliability for polyscores and IRT scores as a function of operational test length for Mathematics Knowledge.

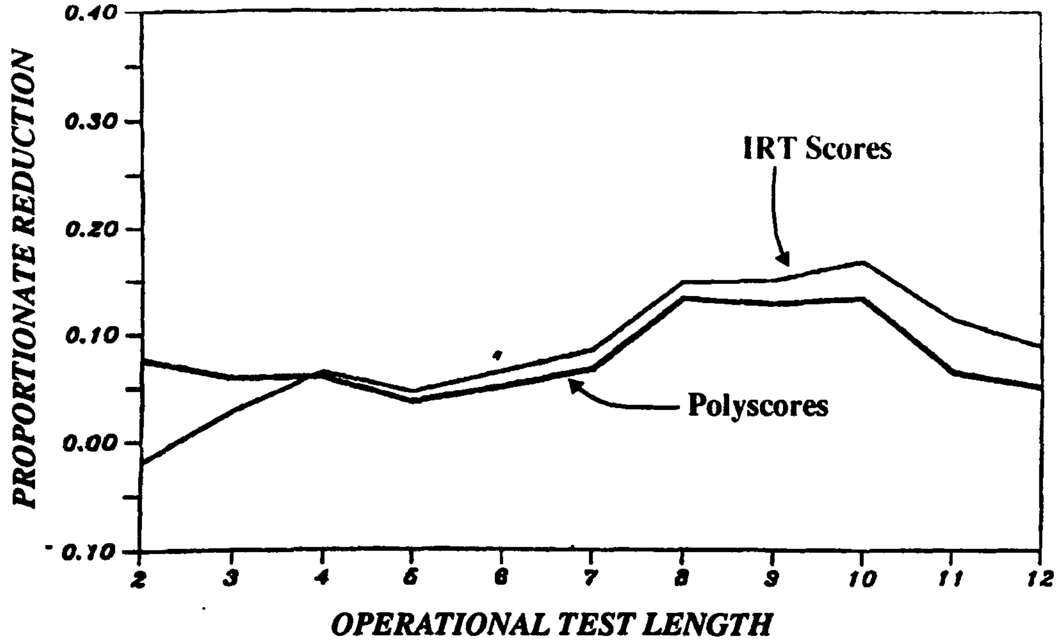

Figure 8. Proportionate reduction in test length under polyweighting and IRT scoring for the operational Mathematics Knowledge forms.

General Science Experimental Tests

Figure 9 is a plot of the median alternate-form correlation for polyscores and for NC scores as a function of test length in the experimental GS alternate forms. Figure 10 compares the median alternate-form correlations obtained for IRT scores with those obtained for NC scores. Figure 11 compares polyscores and IRT scores.

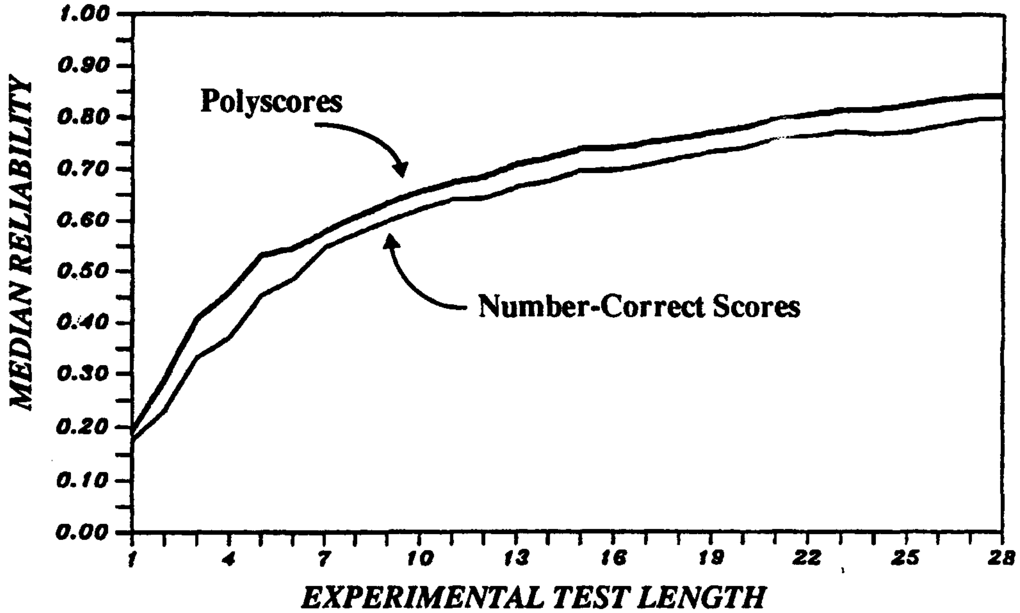

Figure 9. Median reliability for polyscores and number-correct scores as a function of test length, for experimental General Science.

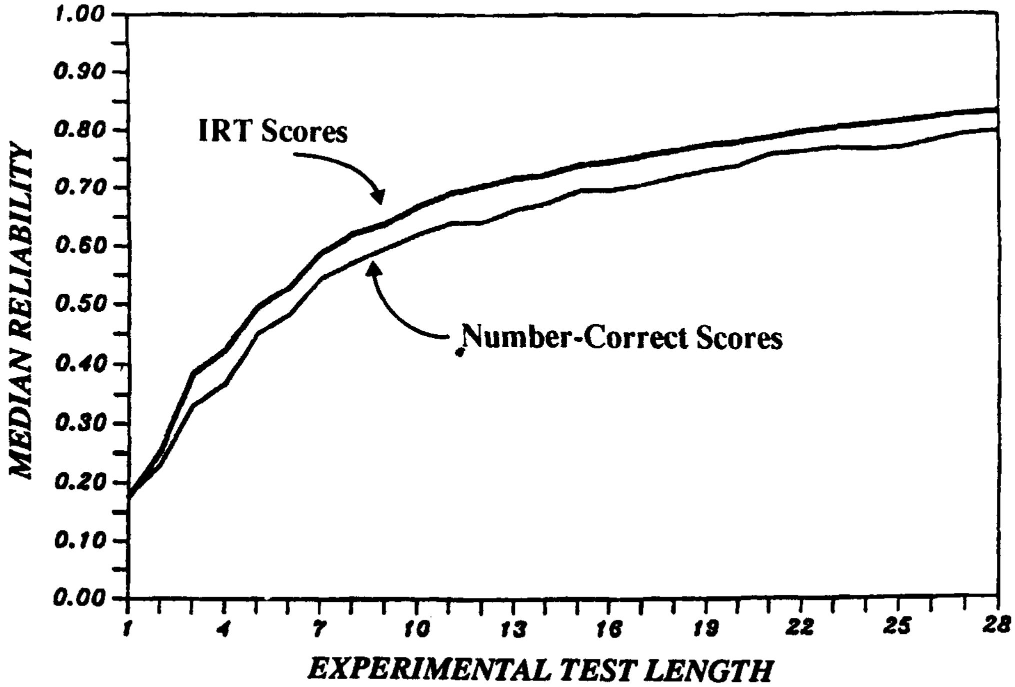

Figure 10. Median reliability for IRT scores and number-correct scores as a function of test length, for experimental General Science.

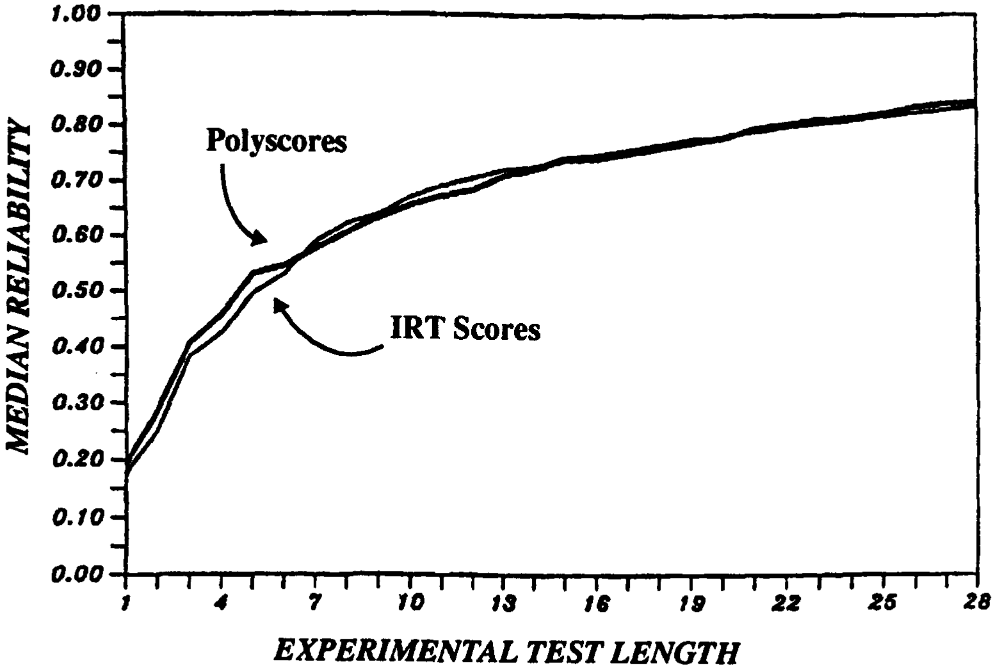

Figure 11. Median reliability for polyscores and IRT scores as a function of test length, for experimental General Science.

Figure 12 shows the proportionate reduction in test length that could be realized, without reducing alternate-form reliability below the median value observed with NC scoring, under polyweighting and under IRT scoring, as a function of test length in the experimental GS alternate forms.

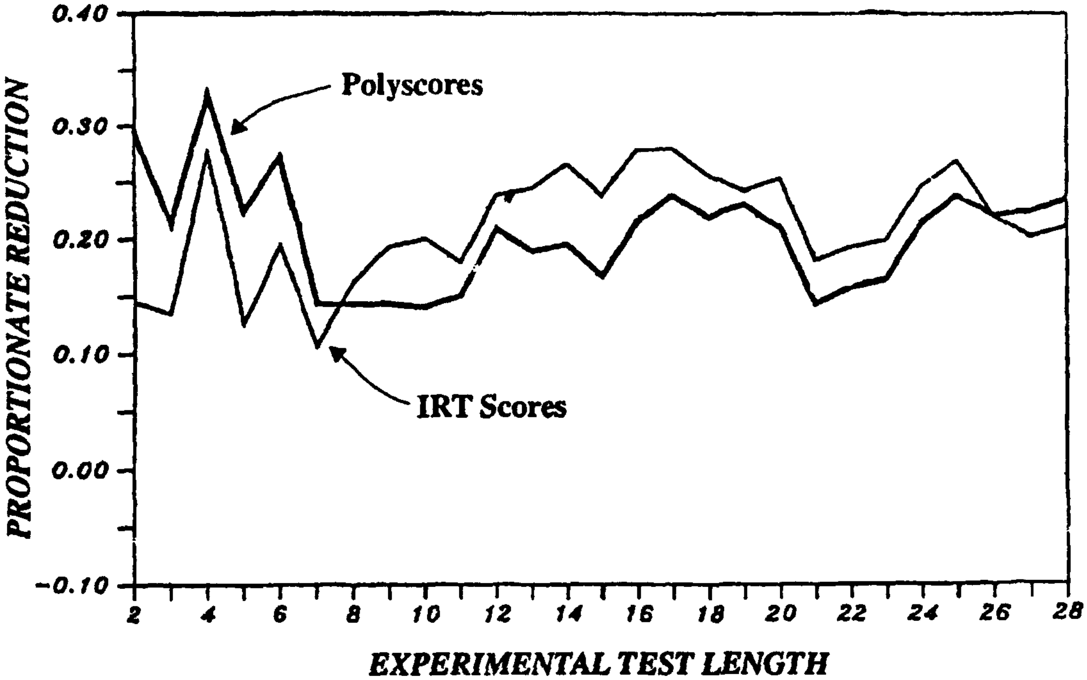

Figure 12. Proportionate reduction in test length under polyweighting and IRT scoring for the experimental General Science forms.

General Science Operational Tests

Figure 13 is a plot of the median alternate-form correlation for polyscores and for NC scores as a function of test length in the operational GS alternate forms. Figure 14 compares the median alternate-form correlations obtained for IRT scores with those obtained for NC scores. Figure 15 compares polyscores and IRT scores.

Figure 16 shows the proportionate reduction in test length that could be realized, without reducing alternate-form reliability below the median value observed with NC scoring, under polyweighting and under IRT scoring, as a function of test length in the operational GS alternate forms.

Discussion

Figures 1 and 2 indicate that polyweighting and IRT scoring both provide higher alternate-form reliability than NC scoring at virtually all experimental MK test lengths. Figure 3 indicates that polyweighting and IRT scoring provide similar levels of alternate-form reliability for these tests, except at test lengths of 4 to 10 items, where there is an advantage in favor of IRT scoring.

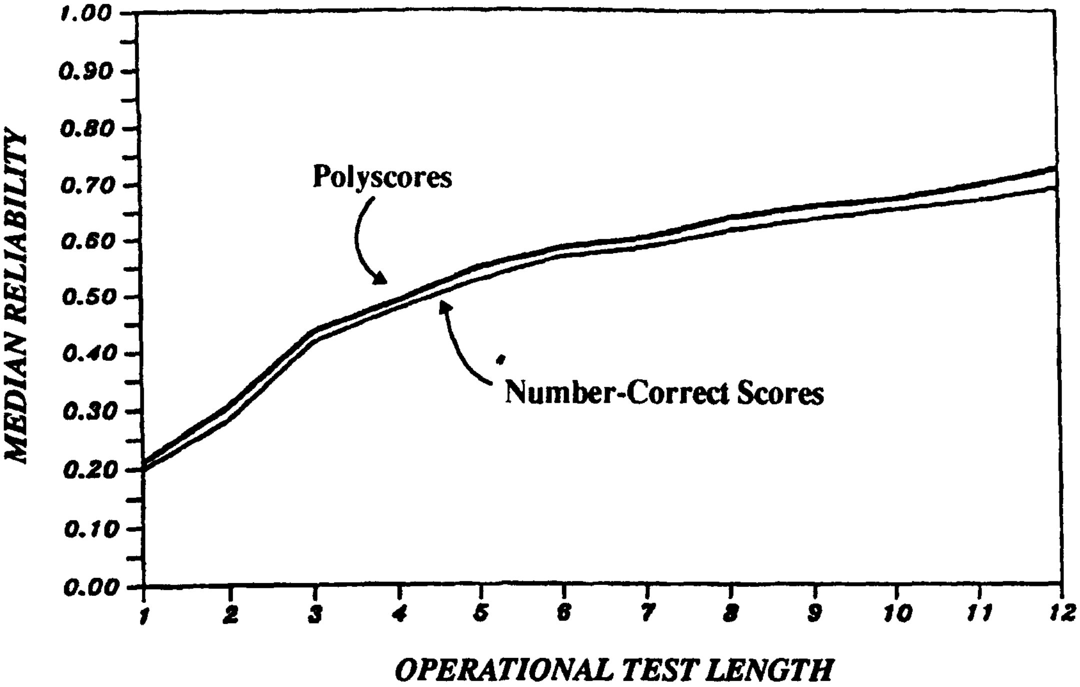

Figure 13. Median reliability for polyscores and number-correct scores as a function of test length, for operational General Science.

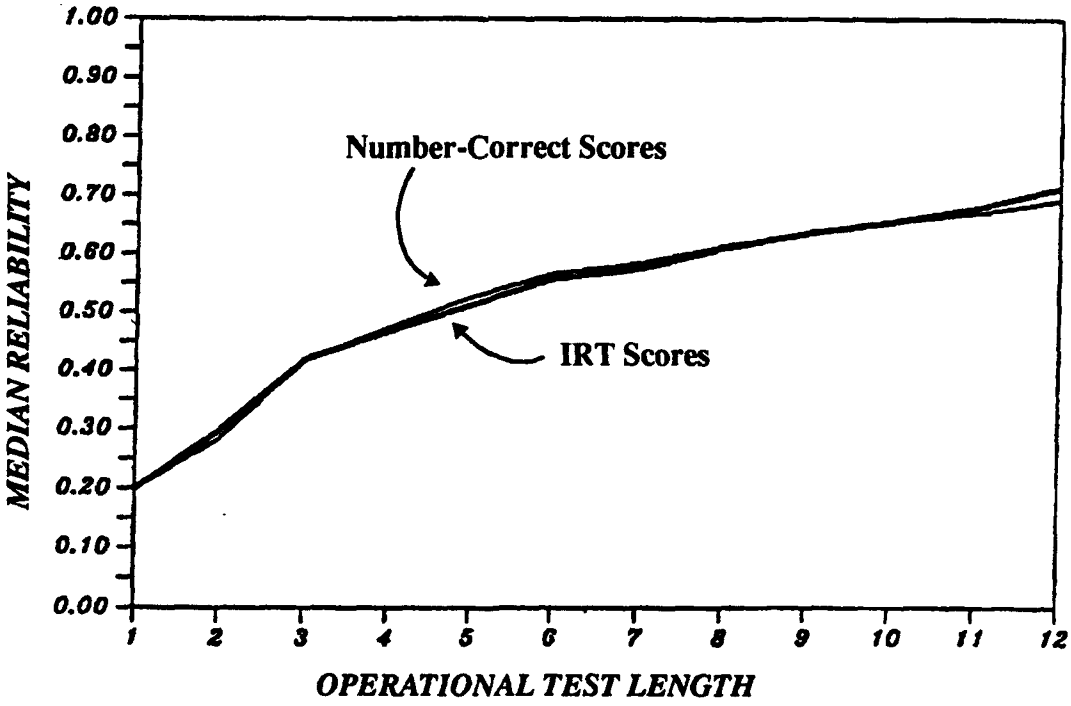

Figure 14. Median reliability for IRT scores and number-correct scores as a function of test length, for operational General Science.

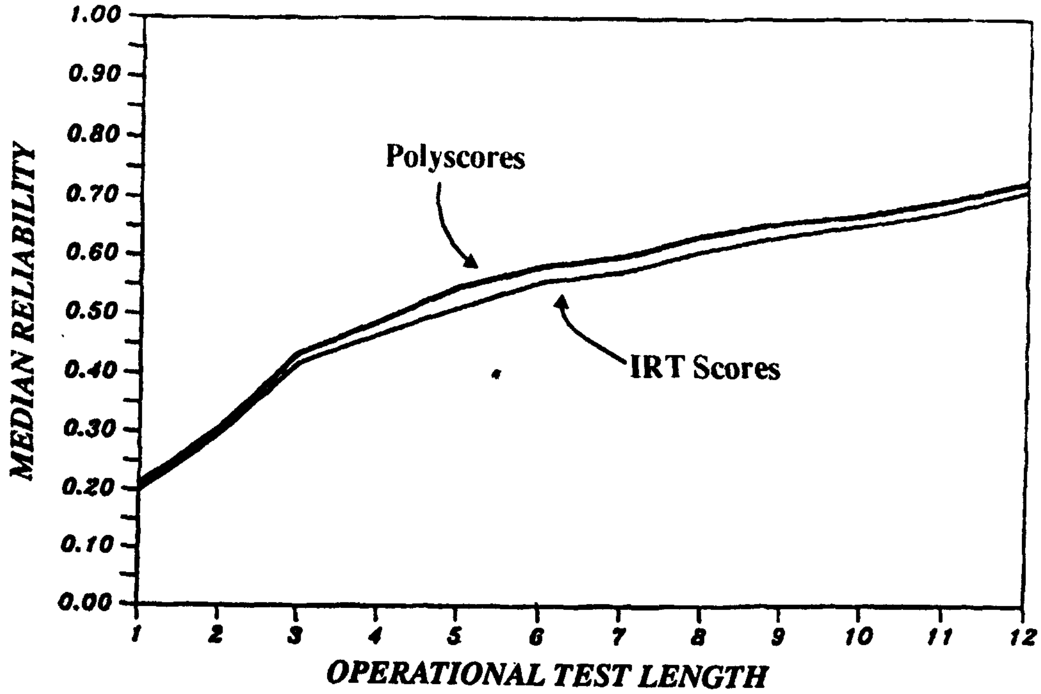

Figure 15. Median reliability for polyscores and IRT scores as a function of test length, for operational General Science.

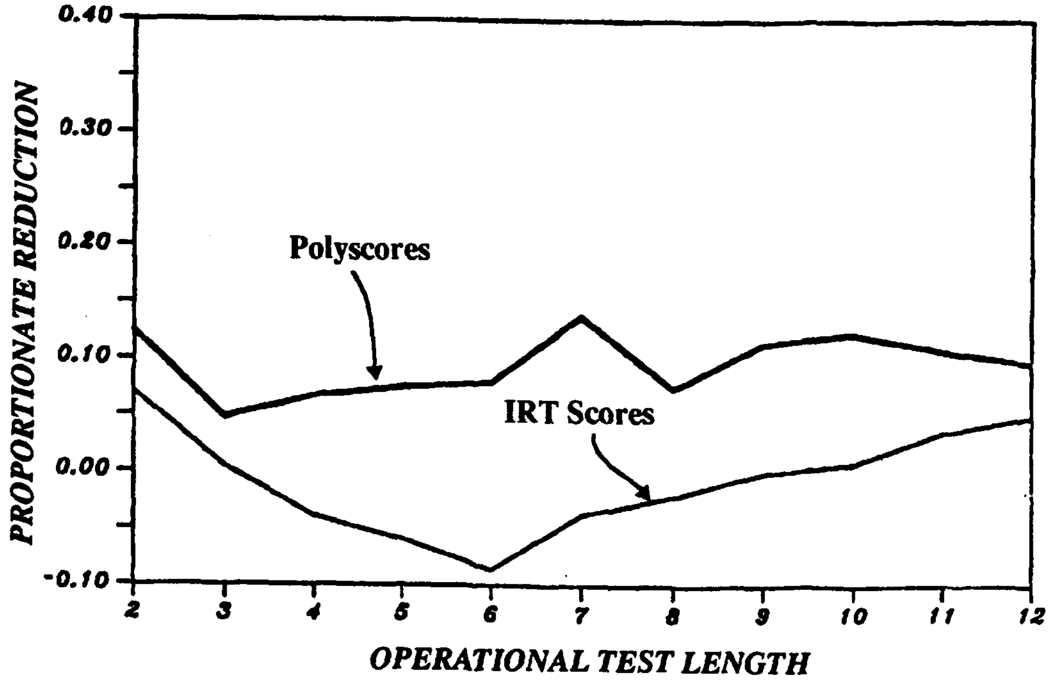

Figure 16. Proportionate reduction in test length under polyweighting and IRT scoring for the operational General Science forms.

Figure 4 suggests that IRT scoring could provide slightly larger reductions in test length than polyscoring at most experimental MK test lengths.The mean reduction in test length, over all test lengths from 2 to 23 items, is 15.3% for IRT scoring and 12.7% for polyweighting, an average superiority of 2.6% for IRT scoring.

Figure 5 indicates that polyweighting provides higher alternate-form reliability than NC scoring at all operational MK test lengths. Figure 6 indicates that IRT scoring is superior to NC scoring at all operational MK test lengths beyond 3 items. Figure 7 indicates that polyweighting is superior to IRT scoring early in the operational MK tests, but that IRT scoring is superior later in these tests.

Figure 8 suggests that IRT scoring could provide slightly larger reductions in test length than polyscoring at operational MK test lengths beyond 4 items. The mean reduction in test length, over all test lengths from 2 to 12 items, is 8.5% for IRT scoring and 7.8% for polyweighting, an average superiority of less than 1% for IRT scoring. However, mean reductions in test length are not very meaningful in cases where the magnitude of the reduction interacts strongly with test length.

Figures 9 and 10 indicate that polyweighting and IRT scoring both provide higher alternate-form reliability than NC scoring at virtually all experimental GS test lengths. Figure 11 indicates that polyweighting and IRT scoring provide similar levels of alternate-form reliability for these tests at test lengths beyond 13 items. At test lengths below 7 items, there is an advantage in favor of polyweighting. For test lengths from 7 to 13 items, there is an advantage in favor of IRT scoring.

Figure 12 suggests that IRT scoring could provide slightly larger reductions in test length than polyscoring at experimental GS test lengths ranging from 8 to 25 items. Below 8 items, polyweighting is superior. The mean reduction in test length, over all test lengths from 2 to 28 items, is 21.2% for IRT scoring and 20.4% for polyweighting, an average superiority of less than 1% for IRT scoring.

Figure 13 indicates that polyweighting provides higher alternate-form reliability than NC scoring at all operational GS test lengths. Figure 14 indicates that IRT scoring is superior to NC scoring very early in the operational GS tests, and also near the end of these tests. However, at test lengths from 3 to 10 items, the median alternate-form correlation for NC scores is as high, or higher, than the median alternate-form correlation for IRT scores. Figure 15 indicates that polyweighting provides higher alternate-form correlations than IRT scoring at all operational GS test lengths.

Figure 16 suggests that polyweighting could provide larger reductions in test length than IRT scoring at all operational GS test lengths. The mean reduction in test length, over all test lengths from 2 to 12 items, is 9.5% for polyweighting and -.7% for IRT scoring, an average superiority of approximately 10% for polyweighting.

The poor performance of IRT scoring when applied to the operational GS tests is surprising in view of the generally good performance of this method on the other tests studied. A diligent search for procedural or computational errors in our analysis of the operational GS data was conducted. None were found. At this point, experimenter error seems unlikely as an explanation, since the same computer programs were used for all analyses in the study and the results obtained for IRT scoring of the operational GS tests are not so bizarre as to be unbelievable.

Following are two possible explanations of the relatively poor performance of IRT scoring with the operational GS tests:

1. IRT item calibration for the 3PL model has been found to work best when there are 1,000 or more examinees per item. In this study, two of the samples available for IRT calibration of operational GS items contained fewer than 800 examinees (see Table 2). Perhaps these samples are too small. (However, an examination of the alternate-form correlations for individual operational GS tests does not suggest a relationship between the magnitude of these correlations and the available sample sizes.)

2. Previous research on the ASVAB has shown the GS test to be multidimensional, with the two primary dimensions corresponding to physical science and life science content. IRT scoring under the 3PL model assumes that the items scored are unidimensional. Perhaps the experimental GS alternate forms, which are more than twice as long as the operational GS alternate forms, are relatively robust against violations of the unidimensionality assumption, while the shorter tests are not as robust.

It is possible that neither of these explanations is correct. Research that identifies the conditions that cause IRT scoring to perform poorly is needed.

Sympson and Haladyna (1993) reported that increases in coefficient-α were smaller for tests that had been assembled using items judged "best" by psychometric criteria. In the current study, a related result for polyscores and also for IRT ability estimates was observed. Under polyweighting, the mean reduction in test length for operational MK alternate forms is smaller than the mean reduction for experimental MK forms (7.8% vs. 12.7%). The mean reduction for operational MK alternate forms is also smaller under IRT scoring (8.5% vs. 15.3%). Similar results are observed for the GS alternate forms. Under polyweighting, the mean reduction in test length for operational GS alternate forms is 9.5%, while the mean reduction observed for experimental GS forms is 20.4%. The mean reduction for operational GS alternate forms is also smaller under IRT scoring (-.7% vs. 21.2%).

The primary difference between the operational and experimental alternate forms created in this study, other than their length, is the fact that the operational items had been selected for inclusion in the ASVAB, while the experimental items were originally prepared for use in a computerized-adaptive test (CAT) item pool. The experimental items intentionally covered a wide range of item difficulties. On the other hand, "best" items, as selected by Sympson and Haladyna (1993) and as selected for the ASVAB, are items with high item-total correlations. Selecting items that have high item-total correlations tends to reduce the range of item difficulties among the items selected.

Whatever the cause, the operational ASVAB items used in this study show less variation in their item difficulties than the experimental items. The standard deviation among IRT difficulty (b) parameters for the operational MK items is .91, while the standard deviation among b parameters for the experimental MK items is 1.16. Similarly, the standard deviation among b parameters for operational GS items is .99, while the standard deviation among b parameters for the experimental GS items is 1.86. It is noteworthy that ordering the four types of tests that were studied in terms of their b-parameter standard deviations also gives the correct ordering of the tests with respect to average test length reduction under both polyweighting and dichotomous IRT scoring (excluding the aberrant outcome for IRT scoring of the operational GS alternate forms).

Conclusions

In both content areas studied, polyweighting provided higher levels of alternate-form reliability than NC scoring. These reliability increases can translate into reductions in test length, if one prefers to save testing time rather than increase test reliability. These results confirm one implication of the earlier study by Sympson and Haladyna (1993).

If not for the results obtained with the operational GS alternate forms, one could conclude that dichotomous IRT scoring and polyweighting provide approximately the same potential for reducing test length, with a small advantage to IRT scoring in the content areas studied. However, IRT scoring performed poorly with the operational GS alternate forms. This should serve as a caution flag until further research clarifies why IRT scoring sometimes works well, and sometimes does not.

It appears that the reduction in test length that is available under either polyweighting or dichotomous IRT scoring is an increasing function of variation in item difficulties. Research that examines this hypothesis is needed.

Results currently available suggest that number/proportion-correct scoring is the least desirable option among the scoring methods studied. The next best option is to use either linear polychotomous scoring (i.e., polyweighting) or dichotomous IRT scoring. Linear polychotomous scoring increases measurement precision by considering the wrong answers an examinee selects. Dichotomous IRT scoring increases measurement precision by computing and using *likelihood functions* (Birnbaum, 1968, p. 455). At present, one cannot predict which of these two strategies for increasing measurement precision will work best in a given context.

Given these conclusions, the most effective scoring strategy available appears to be polychotomous IRT scoring. This scoring strategy increases information about examinee knowledge/ability by considering the wrong answers an examinee selects, and also by computing and using likelihood functions. The one empirical study that has compared alternate-form reliabilities of polychotomous IRT scores and dichotomous IRT scores (Sympson, 1993a) found that polychotomous IRT scores were more reliable. Of course, dichotomous and polychotomous IRT scoring both require large samples (1,000 or more examinees) for item calibration. If one does not have samples of this magnitude available, then polyweighting may be an effective way to increase the measurement precision of tests.

References

Birnbaum, A. (1968). 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* (chapters 17-20). Reading, MA: Addison-Wesley.

Haladyna, T. M., & Sympson, J. B. (1988, April). Empirically-based polychotomous scoring of multiple-choice test items: Historical overview. Talk presented in C. E. Davis (Chair), *New Developments in Polychotomous Item Scoring and Modeling*. Symposium conducted at the annual meeting of the American Educational Research Association, New Orleans, LA.

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.

Prestwood, J. S., Vale, C. D., Massey, R. H., & Welsh, J. R. (1985). *Armed Services Vocational Aptitude Battery: Development of an adaptive item pool* (AFHRL-TR-85-19). Brooks Air Force Base, TX: Manpower and Personnel Division.

Sympson, J. B. (1993a). *Extracting information from wrong answers in computerized adaptive testing* (NPRDC-TN-94-1). San Diego: Navy Personnel Research and Development Center.

Sympson, J. B. (1993b). *A procedure for linear polychotomous scoring of test items* (NPRDC-TN-94-2). San Diego: Navy Personnel Research and Development Center.

Sympson, J. B., & Haladyna, T. M. (1993). *An evaluation of "polyweighting" in domain-referenced testing* (NPRDC-TN-94-3). San Diego: Navy Personnel Research and Development Center.

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

Distribution List

Distribution:
Office of the Assistant Secretary of Defense (FM&P)
Office of Naval Research (Code 1142) (3)
Defense Technical Information Center (DTIC) (12)

Copy to:
Office of Naval Research (Code 20P), (Code 222), (Code 10)
Naval Training Systems Center, Technical Library (5)
Office of Naval Research, London
Director, Naval Reserve Officers Training Corps Division (Code N1)
Chief of Naval Education and Training (L01) (2)
Chief of Naval Operations(N71)
Curriculum and Instructional Standards Office, Fleet Training Center, Norfolk, VA
Director, Recruiting and Retention Programs Division (PERS-23)
Commanding Officer, Sea-Based Weapons and Advanced Tactics School, Pacific
Commanding Officer, Naval Health Sciences Education and Training Command, Bethesda, MD
Marine Corps Research, Development, and Acquisition Command (MCRDAC), Quantico, VA
AISTA (PERI II), ARI
Armstrong Laboratory, Human Resources Directorate (AL/HR), Brooks AFB, TX
Armstrong Laboratory, Human Resources Directorate (AL/HRMIM), Brooks AFB, TX
Armstrong Laboratory AL/HR-DOKL Technical Library, Brooks, AFB, TX
Library, Coast Guard Headquarters
Superintendent, Naval Post Graduate School
Director of Research, U.S. Naval Academy
Naval Education and Training Program (NETPMSA, Code 047), Pensacola (N. N. Perry)