Navy Personnel Research and Development Center
San Diego, California 92152-7250 TN-94-3 October 1993

An Evaluation of "Polyweighting" in Domain-Referenced Testing

J. Bradford Sympson
Thomas M. Haladyna

NPRDC-TN-94-3 October 1993

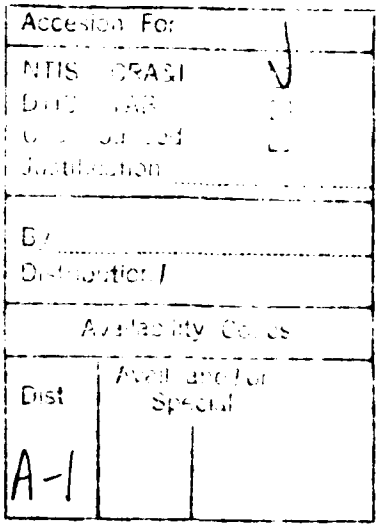
An Evaluation of "Polyweighting" in Domain-Referenced Testing

J. Bradford Sympson

Thomas M. Haladyna
Arizona State University West
Phoenix, Arizona 85069-7100

Reviewed by
Daniel O. Segall

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



Approved for public release;
distribution is unlimited.

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

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

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

4. TITLE AND SUBTITLE
An Evaluation of "Polyweighting" in Domain-Referenced Testing

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

6. AUTHOR(S)
J. Bradford Sympson, Thomas M. Haladyna

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-TN-94-3

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 AGENCY REPORT NUMBER

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 technical note describes an empirical evaluation of a polychotomous item scoring procedure developed by the first author. This new scoring procedure *(polyweighting)* assigns an empirically-derived scoring weight to each possible response to a test question. An examinee's *polyscore* is equal to the mean of the scoring weights of the response categories chosen by the examinee.

In this research, polyweighting was applied to test data obtained from 1,100 resident physicians who had completed a 200-item medical certification test. Using the 200 items as an item bank, the authors assembled 20 short (10-, 20-, 30-, 40-item) assessment tests and used both proportion-correct scores and polyscores from these short tests to predict each physician's score on the 200-item certification test.

For all 20 assessment tests, polyweighting resulted in higher cross-validated internal-consistency reliability (coefficient-α) and domain validity. 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.

14. SUBJECT TERMS	15. NUMBER OF PAGES
Selection, classification, training, testing, item scoring, polychotomous scoring, polytomous scoring	15
	16. PRICE CODE

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

Foreword

This technical note describes an empirical evaluation of a polychotomous item scoring procedure developed by the first author (Sympson). The new procedure should be particularly useful in situations involving medium-sized (N = 100-1,000) item calibration samples and/or multidimensional item content domains.

Results reported in this technical note were originally presented in a symposium titled *New Developments in Polychotomous Item Scoring and Modeling* (C. E. Davis, Chair) at the annual meeting of the American Educational Research Association, which was held in New Orleans in April of 1988. It is being published at this time for archival purposes.

The research described here was conducted under the Navy Personnel Research and Development Center Independent Research and Independent Exploratory Development (IR/IED) Programs. 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). Preparation of this document was funded by the Office of Naval Research (Code 1142) under the Navy Laboratory Participation Program (Program Element 0601153N, Work Order R4204).

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

In this research, the authors conducted an empirical evaluation of a new test scoring procedure (*polyweighting;* Sympson, 1993) in the context of medical certification testing. 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." The computer program POLY was applied to the Sample A data in order to obtain summary statistics and polyweights for all 200 items.

Using the set of 200 items as an item bank, the authors 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, internal-consistency reliability coefficients were computed and both types of test score were correlated with Sample B 200-item domain scores.

Results

For all 20 assessment tests, polyweighting resulted in higher cross-validated internal-consistency 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.

Conclusions

Results of this study indicate that polyweighting can provide consistent increases in test reliability and domain-related validity. These findings also suggest that polyweighting should allow test developers to reduce test length, while maintaining test reliability at the level observed under traditional number/proportion-correct scoring.

Recommendation

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

Contents

Introduction

Polychotomous scoring of multiple-choice test items is based on the assumption that ability (knowledge/skill) distributions are not the same for examinees who choose different response options, even if they have answered the same number of items correctly. If this assumption is correct, additional information about an examinee's knowledge/skill-level can be obtained by noting which questions the examinee has answered correctly and which incorrect answers were selected.

A variety of polychotomous scoring methods have been tried, dating from about 1935 to the present (Haladyna & Sympson, 1988). These methods can be classified as either *linear* or *nonlinear*. Linear polychotomous scoring involves the use of fixed scoring weights that vary over response options. Nonlinear polychotomous scoring is based on item response theory (IRT) and involves the use of *likelihood functions* (Birnbaum, 1968, p. 455). Since realistic IRT models require large sample sizes ($N \geq 1000$) for item calibration, and since test scoring under these models usually requires an assumption that the test is unidimensional, nonlinear polychotomous scoring is less widely applicable than linear polychotomous scoring.

Sympson (1993) has introduced a new method for linear polychotomous scoring called *polyweighting*. The scores obtained with this method are called *polyscores*. The purpose of this study was to compare polyscores with traditional proportion-correct scores in terms of their internal-consistency reliabilities and domain validities. Comparisons are made in a context similar to that found in certification, licensing, proficiency, or competency testing.

Polyweighting

The category scoring weights used in polyweighting are called *polyweights*. An examinee's polyscore is equal to the mean of the polyweights for the categories chosen by the examinee. The iterative procedure used to derive polyweights for a set of items is described in Sympson (1993) and implemented in the computer program POLY. 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

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

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, $\overline{R}_{ij}$ 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. Thus, under polyweighting, examinees never receive more credit for an incorrect answer than for a correct answer.

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. Since polyweights are derived from examinee percentile ranks, and since percentile ranks are independent of the difficulty of the items administered, polyweights obtained for an item are independent of the difficulty of the other items administered.

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.

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

Data from 1,100 physicians who 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." The program POLY was then applied to the Sample A data to obtain item summary statistics and polyweights for all 200 items.

Next, using the set of 200 items as an item bank, 20 different assessment tests were assembled and scored in Sample B. These tests were as follows:

1. Three randomly-selected item-sets of size 10 were designated as tests R10-1, R10-2, and R10-3. Three samples of items were used in order to obtain an indication of the amount of sampling variation in reliability and domain validity that could be expected when tests are assembled by

randomly sampling items. Since items in the 200-item test had been allocated to five content categories by expert (physician) consultants, two items were randomly selected from each content category, in order to ensure that each test was content valid.

2. In a manner similar to the 10-item tests, three randomly-selected item-sets of size 20 were designated as tests R20-1, R20-2, and R20-3. Each of these tests included items from one of the randomly-assembled 10-item tests. R20-1 included the items making up test R10-1, R20-2 included the items in test R10-2, and R20-3 included the items in test R10-3. In these tests, four items were randomly selected from each of the five content categories.

3. Three randomly-selected item-sets of size 30 were designated as tests R30-1, R30-2, and R30-3. Each of these tests included items from one of the randomly-assembled 20-item tests. R30-1 included the items making up test R20-1, R30-2 included the items in test R20-2, and R30-3 included the items in test R20-3. In these tests, six items were randomly selected from each of the five content categories.

4. Three randomly-selected item-sets of size 40 were designated as tests R40-1, R40-2, and R40-3. Each of these tests included the items from one of the randomly-assembled 30-item tests. R40-1 included the items making up test R30-1, R40-2 included the items in test R30-2, and R40-3 included the items in test R30-3. In these tests, eight items were randomly selected from each of the five content categories.

5. Using the results of the Sample A 200-item POLY run, tests of length 10, 20, 30, and 40 items were assembled using "traditional" item selection criteria. In this test construction procedure, items were selected that had the highest correct-answer point-biserial correlations (Henrysson, 1971, p. 142), subject to a requirement that all item difficulties (proportions correct) had to be within .10 of the mean item difficulty in the 200-item domain. The resulting tests were designated as tests T10, T20, T30, and T40. Test T20 included the items making up test T10, test T30 included the items in test T20, and test T40 included the items in test T30. As before, item selection was accomplished within the designated content categories, with k items being selected from each category for a $5k$-item test.

6. Using the results of the Sample A 200-item POLY run, tests of length 10, 20, 30, and 40 items were assembled by selecting the items within each content category that had the highest η coefficients (Lord & Novick, 1968, p. 263). In this context, the squared η coefficient for an item indicates the proportion of variance in percentile ranks that is accounted for by knowing which response category each examinee has selected. These four tests were designated as tests EM10, EM20, EM30, and EM40. Test EM20 included the items making up test EM10, test EM30 included the items in test EM20, and test EM40 included the items in test EM30. As before, k items were selected from each content category for a $5k$-item test.

Each of the 20 tests described above was scored two different ways in Sample B. First, each test was scored by assigning a weight of 1 to all correct-response categories, a weight of 0 to all incorrect-response categories, and computing the mean weight among the categories selected. This gave the traditional proportion-correct (PC) score. Next, each test was scored using the polyweights derived in Sample A. For each Sample B examinee, his/her polyscore was the mean Sample A polyweight among the categories selected by the examinee.

For each of the 20 tests, Sample B item and test scores were used to compute coefficient-α (Cronbach, 1951) for both PC scoring and for polyweighting. The two resulting values of α for each test were then used to compute a value of the following *relative information index*:

$$H = \frac{\alpha_p(1 - \alpha_d)}{\alpha_d(1 - \alpha_p)} \quad . \tag{2}$$

This index is based on the Spearman-Brown formula (Lord & Novick, 1968, p. 112). The Spearman-Brown formula gives the reliability of a lengthened test as a function of the initial reliability of the test and the proportionate increase in test length that is anticipated. However, rather than use the Spearman-Brown formula to predict reliability, one can rearrange the formula and use it to determine how much a given test would have to be increased in length in order to obtain a specified level of reliability (Nishisato, 1980, p. 118).

In Equation 2, α_d is the value of coefficient-α obtained under PC scoring and α_p is the value of coefficient-α obtained under polyweighting. This information index indicates the proportionate increase in test length that would be required in order to achieve the same reliability under PC scoring that was achieved using polyweighting.

Next, for each of the 20 tests, Sample B test scores and Sample B domain scores (based on all 200 items) were used to compute domain validities. For PC scoring, each examinee's domain score was the examinee's proportion correct on the 200-item test. For polyweighting, examinee domain scores were obtained by running POLY on the Sample B data for all 200 items. It is relevant to note that under PC scoring the weight (1 or 0) assigned to any given response category was the same when an item appeared in a short assessment test and when it was part of the domain. On the other hand, as a result of sampling error, the Sample A polyweight assigned to a response category during scoring of an assessment test in Sample B was, in general, somewhat different than the weight assigned to that category during the computation of Sample B domain scores.

Finally, after computing two Sample B domain validities for each test, the difference was obtained for each test:

$$D = \rho_p - \rho_d \quad , \tag{3}$$

where ρ_p is the domain validity under polyweighting and ρ_d is the domain validity under PC scoring.

Results and Discussion

Table 1 shows the results of this comparative evaluation of PC scoring and polyweighting. Inspection of Table 1 shows that for all combinations of test length and test-construction method, polyweighting outperforms PC scoring in the cross-validation sample.

Table 1

Cross-validated Reliability and Domain Validity of Proportion-correct Scores and Polyscores for 20 Tests

	Reliability (α)			Domain Validity		
	Type of Score			Type of Score		
Test	PC	Poly	*H*	PC	Poly	*D*
R10-1	.252	.322	1.41	.272	.376	.104
R10-2	.299	.339	1.20	.288	.369	.081
R10-3	.355	.461	1.56	.437	.538	.101
R20-1	.517	.580	1.29	.531	.635	.104
R20-2	.534	.586	1.24	.560	.623	.063
R20-3	.508	.623	1.60	.577	.705	.128
R30-1	.647	.697	1.26	.690	.757	.067
R30-2	.582	.646	1.31	.634	.695	.061
R30-3	.599	.691	1.50	.658	.764	.106
R40-1	.701	.755	1.31	.758	.826	.068
R40-2	.675	.727	1.28	.731	.786	.055
R40-3	.701	.777	1.49	.755	.828	.073
T10	.583	.605	1.10	.597	.664	.067
T20	.720	.740	1.11	.751	.812	.061
T30	.778	.799	1.14	.815	.870	.055
T40	.824	.841	1.13	.847	.899	.052
EM10	.625	.656	1.15	.606	.673	.067
EM20	.738	.766	1.16	.760	.819	.059
EM30	.810	.833	1.17	.815	.881	.066
EM40	.843	.862	1.17	.841	.911	.070

As expected, both coefficient-α and domain validity increase as test length increases, regardless of test-construction method and scoring method. Also, as might be expected, both

coefficient-α and domain validity are higher for the systematically-constructed tests than for the randomly-assembled tests.

For each test length, tests made up of items with maximum η coefficients are more reliable than tests assembled using the traditional method. However, under PC scoring the "EM" tests are not always superior to the "T" tests when domain validity is the criterion.

For the randomly-assembled tests (R10-1 through R40-3), the *H* statistics in column 4 indicate that, on the average, polyweighting increased coefficient-α by an amount that corresponds to a 37% increase in test length. Smaller increases are observed for the systematically-constructed tests, where the mean value of *H* is 1.14. There is an indication that the EM tests benefit slightly more from polyweighting, since the mean *H* for these four tests is 1.16, vs. 1.12 for the four T tests.

The *D* statistics in column 7 indicate that, on the average, polyweighting increased domain validity for the randomly-assembled tests by .084. For the traditionally-constructed (T) tests, the mean value of *D* is .059. For the EM tests, the average increase in domain validity is .066. Over all 20 tests, the minimum increase in domain validity is .052.

An important comparison that is implicit in Table 1 can be obtained by contrasting α-coefficients and domain validities of tests that were assembled using the traditional method and scored dichotomously with those of tests that were assembled using η-coefficients and scored polychotomously. This provides a comparison between currently prevailing (dichotomous) test-construction and scoring practice and an alternative (polychotomous) approach. Comparison of α-coefficients (.656 vs. .583, .766 vs. .720, .833 vs. .778, and .862 vs. .824) results in a mean *H* statistic of 1.35, indicating that a combination of polychotomous item-selection and scoring provides an increase in reliability that corresponds to a 35% increase in test length. Comparison of domain validities (.673 vs. .597, .819 vs. .751, etc.) results in a mean *D* statistic of .069, with a minimum increase in domain validity of .064.

Conclusion

Results of this study indicate that polyweighting can provide consistent increases in test reliability and domain-related validity. The findings also suggest that polyweighting should allow test developers to reduce test length, while maintaining test reliability at the level observed under traditional number/proportion-correct scoring.

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.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

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.

Henrysson, S. (1971). Gathering, analyzing, and using data on test items. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 130-159). Washington, DC: American Council on Education.

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

Nishisato, S. (1980). *Analysis of categorical data: Dual scaling and its applications*. Toronto, Canada: University of Toronto Press.

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

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)
Curriculum and Instructional Standards Office, Fleet Training Center, Norfolk, VA
Chief of Naval Operations (N71)
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)