Gorman, Steven, Annandale, Virginia. (Wed. A.M.)

Computerized Optimal Weight Scoring (COWS): A Comparison of Three Procedures

This paper documents the results of a simulation study comparing three computerized optimal weight scoring (COWS) procedures based on item response theory, with conventional scoring. The effect of the COWS procedures is to differentially weight test items as a function of examinee ability and the item characteristics. Low ability examinees are given very low weights on difficult test items, lowering the effects of guessing, and decreasing test error.

A MONTE CARLO COMPARISON OF THREE OPTIMAL TEST SCORING PROCEDURES

Steven Gorman
Electronic Data Systems, alex, va

Differentially weighting items to improve the psychometric properties of multiple-choice aptitude or achievement test scores has long been of theoretical and practical concern. Applications of item weighting procedures which were empirically developed or based on classical test theory did not generate the desired psychometric properties (Bejar & Weiss, 1973; Downey, 1976).

Lord (1968) successfully applied item weighting based on the three parameter logistic model developed by Birnbaum (1968) to the Verbal Scholastic Aptitude Test (VSAT). The effect of this procedure is to differentially weight test items as a function of examinee ability on the basis of their item discriminatory power, difficulty, and susceptibility to guessing. Accordingly, low ability examinees will be given very low weights on difficult test items, thus lowering the effects of guessing, and decreasing test error. The use of optimal item weights eliminates more test score error at low ability ranges, where guessing on difficult items is more prevalent. Because of the difficulty (until recently) of accurately estimating these item parameters, and the computations involved in using Lord's procedure, this optimal item weighting method has not been widely adopted. The accurate estimation of the three parameters of test items has become commonplace with the development of several new computer programs, namely LOGIST (Wood, Wingersky, & Lord, 1976), ANCILLES (Croll & Urry, in preparation) and FITZIP (Gugel, in preparation). Three FORTRAN computer programs, MAXLIKE, BAZEMODAL, and OWENSTAT have been developed to estimate ability based on the use of optimal scoring weights.

The present study simulates a live testing situation by introducing item parameter estimation error. It compares the performance of conventional unit weight scoring with three optimal item weighting methods using the item parameters estimated by ANCILLES. The data in this investigation were produced with a Monte Carlo procedure which generated examinee item responses.

TECHNIQUES INVESTIGATED

The psychometric characteristics of test scores for three optimal item weight scoring procedures, and a unit weighting scoring method were investigated using three idealized types of test distributions.

The three optimal scoring procedures are based on Birnbaum's three parameter logistic model which states that the probability of a correct response given an ability level is:

$$P(u_i = 1 \mid \theta) = c_i + (1 - c_i)[1 + \exp(-1.7a_i(\theta - b_i))]^{-1} \qquad (1)$$

where a_i is the item discriminatory power, b_i is the item difficulty, and c_i is the item coefficient of guessing.

The three optimal scoring procedures, and a conventional scoring procedure used in the study are described in the following paragraphs.

Owen Bayesian Scoring

This method is described in Gorman (1979) and based on Owen (1975). The method uses the Owen algorithm scoring procedure in a sequential rather than adaptive mode, in the computer program OWENSTAT.

Bayes Modal Scoring

This procedure is based on the work on Samejima (1969) and assumes that ability is normally distributed. The ability estimate is the maximum value for:

$$B_v(\theta) = N(\theta) \prod_{i=1}^{n} P^u i(\theta) Q^u i(\theta) \qquad (2)$$

where $N(\theta)$ is the normal Gaussian distribution, $P^u i(\theta)$ is the probability of a correct response to item i given ability θ, and $Q^u i(\theta) = 1 - P^u i(\theta)$.

Maximum Likelihood Scoring

This procedure, described in Lord & Novick (1968), states that the ability estimate is that value which maximizes $L_v(\theta)$ in:

$$L_v(\theta) = \prod_{i=1}^{n} P^u i(\theta) Q^u i(\theta) \qquad (3)$$

where $L_v(\theta)$ is the maximum likelihood ability estimate.

The above two scoring equations were solved in the computer programs BAZEMODAL and MAXLIKE by use of a modified Newton-Raphson algorithm.

Z-Score Transformation

This is a conventional unit weighting scoring procedure where:

$$Z = \frac{X - \bar{X}}{S.D.} \qquad (4)$$

where X is the examinee's raw test score, $\bar{X}$ is the average raw test score, S.D. is the standard deviation of raw scores, and Z is the ability estimate.

METHODOLOGY

Ideal tests were constructed consisting of two levels of item quality, low ($a_i = 0.8$) and high ($a_i = 1.6$). Three test types were developed which varied the item difficulty (b_i) distributions in order to provide maximum test information (Birnbaum, 1968) at:

(1) evenly distributed values over the ability range $\theta = -2.5$ to $+2.5$ (rectangular)

(2) midpoints of even-sized areas of the Gaussian distribution (normal) and

(3) at the ability mean (peaked).

The susceptibility to guessing, c_i was set for all test items at .15, a reasonable average for a five alternative multiple choice test (based on Jensema, 1976).

Data were generated in accordance with Birnbaum's (1968) three parameter logistic model (equation 1) in two parts:

(1) creating examinee responses for the estimation of item parameters, and
(2) creating examinee responses to all test items with known item parameters.

When the item parameters are specified, the probability of a correct response is strictly a function of the simulated examinee's (sims) ability. To generate dichotomous responses, the probability of a correct response is determined for the examinee's ability from equation 1. If this probability is greater than or equal to a random number between 0 and 1, then a correct response is generated, else an incorrect response is generated. The item parameters for these items along with items from a separate study (Gorman, 1980) were then estimated by ANCILLES based upon the administration of tests of 51 item length to 2000 sims representing a normal population. These estimated item parameters were used in the three optimal scoring procedures.

STUDY 1

In Study 1, a group of 500 normally distributed sims, generated by the LIRANDOM computer program (Learmonth & Lewis, 1973), were administered all tests. The criterion evaluated in Study 1 was the fidelity coefficient, the correlation of known ability with ability estimated by each scoring method on the 20 and 30 item tests. This statistic has been widely used in other testing research. Samejima (1977) decries the use of this statistic, stating that since it is a correlation coefficient, it is dependent not only upon the test, but also upon the specific group of examinees tested. Testing the same sims from a normally distributed population will hold constant across test scoring methods the effect of the ability distribution of examinees. It should be demonstrable that a theoretical test can have a high fidelity coefficient, yet have poorer psychometric properties as a function of ability than other theoretical tests.

RESULTS

The results of Study 1 are displayed in Table 1. The fidelity coefficients, the correlations between known and estimated ability, are typically greater for tests of all types, lengths, and item qualities scored with the optimal scoring methods. The only exception is with the peaked high item quality tests, where the conventional scoring procedure ordered examinees better, on average, than with the maximum likelihood procedure. On the low item quality tests, the ordering of examinees is greatest with the peaked tests. On the high item quality test, the normal tests ordered examinees best. Note that the 20 item normal test with higher item quality scored with any of the optimal scoring procedures has higher fidelity coefficients than the 30 item conventionally scored test. Also note that as test length and item quality increase, the fidelity coefficients of the optimal scoring procedures become much greater than those of the conventional method.

TABLE 1

Correlations between Known Ability and Ability Estimated by Four Scoring Methods on Three Types of Test Distributions*

	20 ITEM TEST					
	Low a_i			High a_i		
Test Type	R	P	N	R	P	N
Method						
BAZEMODAL	842	891	877	927	906	943
MAXLIKE	839	888	874	926	876	941
OWENSTAT	840	892	875	922	911	939
Z-SCORE	828	888	869	908	904	929
	30 ITEM TEST					
BAZEMODAL	907	923	923	949	926	954
MAXLIKE	905	921	925	949	891	954
OWENSTAT	905	924	922	944	924	949
Z-SCORE	894	919	918	933	912	938

*Decimals omitted

DISCUSSION

With low item quality, the 20 item peaked test and the 30 item peaked and normal tests provided the greatest fidelity coefficients, the correlation between the ability estimates and known ability. With high item quality, the normal tests have the greatest fidelity coefficient. This follows from item response theory, which holds that item information becomes more leptokurtic as item discriminatory power (a_i) increases. Thus, since item information is additive, a peaked test with low a_i values should differentiate examinees over a much broader range than a peaked test with high a_i values.

This study has shown that on this one criterion, the non-rectangular tests are better average measures of ability. Study 2 will examine two other psychometric properties of these test scoring methods.

STUDY 2

In Study 2, the sample consisted of 100 sims at each of 11 evenly spaced ability values on the ability continuum -2.5 to +2.5. The instruments used were the three types of 30 item tests. These sims provided data to compute statistics as a function of examinee ability. The criteria evaluated are score bias and test score precision. Test score bias is the average difference between the known examinee ability and the ability estimated by each scoring method. Test score precision is given by the test score information value, an indicator of the usefulness of the test scores for differentiating ability at that ability level. Test score information is inversely related to the square of the standard error of the ability estimate, and varies as a function of ability.

RESULTS

The mean difference between known ability and the ability estimated by the four procedures was computed for the 100 sims at each of the 11 ability levels. For ease of comparison, these data were aggregated by computing an average of the absolute values of the 11 score bias figures. These values are located in Table 2. The data show that the bias is lower (excepting the Bayes modal scored peaked test of low item quality) for all optimal scoring methods for all test types and levels of item quality than with conventional scoring. The bias is most severe with the peaked test, and this bias increases with higher item discriminatory power.

TABLE 2

Average Absolute Value of Score Bias for Ability Estimated by Four Scoring Methods on Three Types of Test Distributions*

	Low a_i			High a_i		
Test Type	R	P	N	R	P	N
Method						
BAZEMODAL	18	27	18	13	38	14
MAXLIKE	11	17	09	12	25	12
OWENSTAT	11	17	10	15	30	14
Z-SCORE	20	26	21	13	44	22

*Decimals omitted

Two criteria need to be measured when evaluating test score precision. One is the average precision (computed by estimated test score information), and the other is the equiprecision over the ability range. Since the items for all types of test distributions have the same item discriminatory power and susceptibility to guessing, all tests share identical potential test score information. The test score information could only be measured over the range -2.0 to +2.0 with this research design, thus some of the test score information for the non-peaked tests is not measured. Therefore, average precision, listed in Table 3, should only be reviewed within test types. For the tests consisting of items with low item discriminatory power, all scoring procedures yielded roughly the same average information. However, with the tests of higher quality items, the optimal scoring procedures provided greater information than with conventional scoring, except with the peaked test.

TABLE 3

Average Test Score Information for Four Scoring Methods on Three Types of Test Distributions

	Low a_i			High a_i		
Test Type	R	P	N	R	P	N
Method						
BAZEMODAL	4.47	5.64	4.99	9.85	10.72	12.47
MAXLIKE	4.39	5.60	4.83	9.62	11.04	12.61
OWENSTAT	4.49	5.57	4.97	9.32	10.45	12.08
Z-SCORE	4.16	5.79	4.86	7.44	12.08	10.25

The equiprecision of the tests by scoring method is measured by the coefficient of variation (CV) of test score information, listed in Table 4. The greater this CV value, the less equiprecise the test score information. The peaked tests provide the greatest CV values, with the optimal scoring procedures yielding greater values than with conventional scoring. The peakedness of information increases, as expected, with the higher item discriminatory power items. The rectangular tests give more even test score precision, with this phenomenon more evident on the higher item discriminatory power test. Although the conventional scoring of the rectangular high item quality test has lower CV values than with optimal scoring, the three optimal scoring methods all yielded greater information at all nine ability levels than with conventional scoring.

TABLE 4

Coefficient of Variation of Test Score Information for Four Scoring Methods on Three Types of Test Distributions

	Low a_i			High a_i		
Test Type	R	P	N	R	P	N
Method						
BAZEMODAL	24	59	25	29	123	53
MAXLIKE	25	63	28	18	136	53
OWENSTAT	27	60	24	22	128	49
Z-SCORE	29	52	36	16	94	52

DISCUSSION

These optimal scoring procedures weight items as a function of their item information (Birnbaum, 1968), the contribution of each item to decrease test score error at each ability level. The item information is a function of the item's a_i, b_i, and c_i parameters. Since the a_i and c_i values are fixed in each test, the full capacity of these scoring procedures is not being demonstrated. Both studies only show the capacity of these procedures to effectively weight items as a function of their appropriateness in difficulty relative to the examinee's ability, not their capacity to weight items as item discriminatory power and item coefficient of guessing vary. In spite of this shortcoming, the optimal scoring procedures show a significant increase in their ability to successfully order the ability of examinees relative to the conventional scoring procedure. The optimal scoring procedures also measure examinees with more precision and less bias than the conventional means.

This study also assumes that the multiple choice test has five response alternatives. For tests with only four item choices, or where the chance of successful guessing is greater that .15, the scoring properties would tend to diminish. The optimal scoring procedure properties would likely diminish slightly, while the conventional scoring properties would drop more significantly. This is due to the optimal weighting procedures capacity to diminish the effect of test score error due to guessing, while conventional scoring procedures are less capable of reducing the effects of guessing.

CONCLUSION

In this Monte Carlo study, three optimal test scoring procedures using estimated item parameters provided better psychometric properties of ability estimates than the conventional procedure. The optimal scoring procedures provided the greatest advantage over conventional scoring on rectangular tests composed of items with high discriminatory power. Test publishers should seriously consider using one of these optimal methods to score their multiple choice examinations. With the ready availability of computer programs to estimate item parameters and optimally score tests, the benefits of enhanced measurement of ability should outweigh the slight increase in computer costs. This study also showed that the fidelity coefficient criterion is only a group psychometric indicator, and does not show the capacity of the test to measure low and high ability examinees. This criterion should be used with caution to avoid making erroneous conclusions.

REFERENCES

Bejar, I.I. and Weiss, D.J. Comparison of four empirical differential item scoring procedures. Research Report 73-2, Psychometric Methods Program, Department of Psychology, University of Minnesota, 1973.

Birnbaum, A. Part 5. Some latent trait models and their use in inferring an examinee's ability. In Lord, F.M. and Novick, M.R. Statistical theories of mental test scores. Reading, Massachusetts: Addison-Wesley, 1968.

Croll, P. and Urry, V.W. ANCILLES: A program for estimation of the item parameters of normal ogive and logistic mental test models.

Downey, R.G. Item option weighting of achievement tests: a comparative study of methods. Paper presented at the American Psychological Association Convention, Washington, D.C., Sept., 1976.

Gorman, S. A comparison of Bayesian adaptive and static tests using a correction for regression. Paper presented at the third Biennial Conference on Computerized Adaptive Testing, Wayzata, Minnesota, 1979.

Gorman, S. A comparative evaluation of two Bayesian adaptive ability estimation procedures with a conventional test strategy. Doctoral dissertation. Catholic University, 1980.

Gugel, J. Estimating the parameters of the normal ogive three parameter model via an extension of theory and methodology of Kelly and of Lord. (in preparation).

Jensema, C.J. A simple technique for estimating latent trait mental test parameters. Educational and Psychological Measurement. 36, 705-715, 1976.

Learmonth, G.E. and Lewis, P.A.W. Naval Postgraduate School Random Number Generator Package: LLRANDOM. Research Report NPS55LW73061A, Naval Postgraduate School, Monterey, California, 1973.

Lord, F.M. An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. Educational and Psychological Measurement, 28, 989-1020, 1968.

Samejima, F. Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph No. 17, 1969.

Samejima, F. A use of the information function in tailored testing. Applied Psychological Measurement, 1977, 1, pp. 233-247.

Wood, R.L., Wingersky, M.S., and Lord, F.M. LOGIST: A computer program for estimating ability and item characteristic curve parameters. Research Memorandum 76-6, Princeton, N.J. Educational Testing Service, 1976.