The Use of Multiple Regression in Evaluating Alternative Methods of Scoring Multiple Choice Tests

Gerald J. Blumenfeld
Isadore Newman
The University of Akron

Echternacht (1972) has reviewed a substantial body of literature in the field of confidence testing. Confidence testing refers to methods of weighing responses so as to reflect the examinee's belief in the correctness of the options selected. The intent is to maximize the amount of information gained from a given set of test items. Lord and Novick (1968) state that maximizing this information involves the manner in which the examinees respond to the items, specifying an item scoring rule, and combining items scores into a weighted total score.

Coombs, Milholland, and Womer (1956) and Ebel (1965) report higher reliabilities for the confidence testing methods they employed when compared to traditional scoring procedures. Echternacht's review (1972)suggests that while higher reliabilities have been found, some researchers have reported lower reliabilities (Hambleton, Roberts, and Traub, 1970; Jacobs, 1971; and Koehler, 1971).

In most studies only increase in reliability has been used to evaluate confidence testing. Minimal attention has been

Presented at the American Psychological Association Convention, at Montreal Canada, August, 1973

given to validity. Archer (1962) has reported lower validity while Hambleton, Roberts, and Traub (1970) have reported higher validity. The purpose of this paper is to provide specific examples of how multiple regression analysis could be used to analyze item discrimination, item validity, and test validity when confidence testing is employed. Current practices tend to utilize apriori scoring formulas rather than maximize the predictiveness possible with the obtained data. We will also suggest that the application of these methods may require the development of multivariate techniques for assessing test reliability.

Method: Data Collection

Subjects and Measures. During the spring quarter, 1973, two sections, 40 students per section, of one of the author's undergraduate test and measurements classes were used to collect the data reported. Students were required to pass 25 M-C item exams covering objectives from each of 6 instructional modules. Each module included initial and remedial exams. A score of 80 percent correct was required. A teaching project was also required, and two of the assignments associated with that project were used as independent criteria for estimates of validity. Only the initial exam of the first three modules was used.

Modules 1, 2, and 3 involved a) types of tests and classification of educational objectives, b) objective test items, and c) anecdotal records, rating scales, and check lists

(including the analytical scoring of essays), respectively. The two assignments used as criteria for assessing validity were 1) the precise statement of a "higher-than-knowledge" behavioral objective; and 2) a three-column table containing a) a higher-than-knowledge behavioral objective, b) a description of an instructional procedure appropriate for the objective, and c) a measurement device which agreed with both the objective and the specific instruction proposed.

Success in developing such a three-column table is one of the major objectives of the course. Therefore, use of these project scores as a criterion for assessing the validity of the exams is appropriate.

Scoring Procedure. Students were required to respond to each four- or five-option multiple choice item twice. They indicated the option they thought least likely to be correct. If the correct option was selected as most likely to be correct, the item was scored two points; if the correct option was selected as least likely to be correct, the item was scored zero points; if the correct option was neither selected as most likely correct nor least likely correct, the item was scored one point.

The statement of a behavioral objective was scored on a zero to five point scale. The objective had to be stated in behavioral terms to receive at least one point. Inclusion of stimulus conditions and required standard of excellence added one point each. If the objective was at the higher-than-knowledge level, this received one point and the omission of

any reference to instruction received one point.

The three-column table was scored on a zero to three point scale. The objective had to describe a higher-than-knowledge level behavior or task to receive at least one point. If the proposed instruction agreed with the objective, a second point was awarded. If the measurement procedure and device agreed with both the objective and the instructional procedure, a third point was awarded.

The authors scored the objectives and the three-column tables independently. Discrepancies were discussed until a common score could be agreed upon. The independent scoring resulted in agreement on more than 80 percent of the papers. Discussion was needed on the other 20 percent.

Results and Discussion

Validity estimates were calculated on two separate criteria. The first criterion was objectives that the students wrote which received grades ranging from 0 through 5. The second criterion for validity estimates was the students project score. This project consisted of writing a behavioral objective, describing how the objective would be taught and how it would be tested. (See method section for more details).

Validity estimates for each of the two criteria were calculated four different ways. These four methods were applied to each of the three tests. The first method (the

traditional method) simply correlated (r) the subject's total score on each test separately with the score they received on criterion one (objectives). Under this condition, the test scores were arrived by traditional grading. Each item was graded either 1 if correct, 0 otherwise.

The second method was identical to the first except in this case each test item was graded in the experimental manner so that the subject could receive for any one item either 0, 1, or 2 points. (See method section for further details). Here, as in the first method, r was used to obtain an estimate of the predictive validity.

The third method used a multiple linear regression procedure to estimate the predictive validity for the experimental procedure. This method differed from the second in that in the second method, each student received only one total score for each of the tests. This score was arrived at by summing the total points earned on each test, separately. In the third method, instead of having one predictor variable, the total test score, three predictor variables were constructed by taking a frequency count of the number of questions each student received full credit (2 points) for, the number of questions on which each received partial credit (1 point), and the number of questions on which each received no credit (0 points). In this manner, information was collected on how many items on each test each student received full, partial, or no credit for. This information then was utilized in the following equation:

Model 1 $Y_1 = a_0U + a_1X_1 + a_2X_2 + a_3X_3 + E_1$

Where Y_1 = the score received on the objectives

X_1 = the number of 0's each student received

X_2 = the number of 1's each student received

X_3 = the number of 2's each student received

U = 1 if the subject is in the sample, 0 if otherwise

a_0, a_1, X_3 = partial regression weights

E_1 = error vector (Y_1-Y_1)

Method four was exactly the same as the third method except that a correction for shrinkage was calculated for the multiple regression formula. The shrinkage formula used was:

$$R^2s = 1 - (1-R^2)\frac{N-1}{N-K}$$

Where: R^2s = the corrected shrunken R^2

R^2 = the calculated R^2

N = the number of independent observations

K = the number of predictor variables

Methods one through four were duplicated exactly using as the criterion, scores on the project in place of scores obtained on the objectives. These results are presented in Tables 1 and 2.

Inspection of Tables 1 and 2 indicates that method two produced a higher predictive validity estimate than did method one, four out of six times. (This was found not to be significant as a Sign Test was used). Method three, the employment of the multiple regression technique, was found to produce higher predictive validity estimates than both methods one and two, six out of six times. This was considered significant since the probability of the Sign Test was $p = .0156$. Method four, in which the R was corrected for shrinkage, was also found to produce higher predictive validity estimates than method one, six out of six times ($p = .0156$) and higher validity estimates than method two, five out of six times ($p = .0938$). This was found to be non-significant at alpha = .05. However, one should keep in mind that the Sign Test is highly conservative.

Seventy-five additional analyses were computed in which each item (25 items per test, on three tests) was used as the predictor variable, predicting the scores on the objectives using methods one and three (traditional scoring and experimental scoring 0, 1, or 2, respectively). Another seventy-five analyses were computed exactly the same way predicting the project score. The results of these analyses can be found in Appendix A. They were not presented in the body of the paper because Tables 1 and 2 are conceptually a composite of all of the separate analyses which are of most theoretical and practical importance.

In addition to estimating the validity of the experiment grading procedures compared to the traditional procedure in predicting the two criteria (objectve and project scores), item discriminations were calculated for each of the items on each of the three tests, comparing both the traditional and experimental grading.

Item discrimination for the traditional method was calculated by correlating (r) the score on each item (graded 1 or 0 with the total score on the test graded in the traditional manner. Therefore, there were twenty-five item discrimination estimates for each of the three tests.

Item discrimination was calculated for the experimental method by using multiple regression analysis to predict the total score for each separate test. These total scores were arrived at by using the experimental grading system (0, 1, or 2 points) and summing these scores for all items to get the total for each test. The predictor variables (the experimental score, or 0, 1, or 2 for each item, was placed into one of three vectors as shown in Model 2.

Model 2: $Y_2 = a_0U + a_1X_4 + a_2X_5 + a_3X_6 + E_2$

Where: Y_2 = the total score for Test 1 using the experimental grading procedure

X_4 = 1 if the subject received no points for item #1 on Test 1, 0 otherwise

X_5 = 1 if the subject received one point for item #1 on Test 1, 0 otherwise

X_6 = 1 if the subject received two poinst for item #1 on Test 1, 0 otherwise

U = 1 if the subject was in the sample, 0 otherwise

l_0, a_1, a_2, a_3 = partial regression weights

E_2 = error vector ($Y_2 - Y_2$)

Seventy-five such models were calculated, one for each of the twenty-five items on each of the three tests.

The results of the item discrimination analyses calculated for both the traditional and experimental grading systems are presented in Tables 3, 4, and 5. Table 3 presents the item discriminations for the twenty-five items in Test 1. As can be seen, when comparing these methods, the experimental method produced higher absolute item discrimination values fifteen out of the twenty-five items on Test 1 (Sign Test not significant).

Table 4 presents the item discriminations for Test 2. Here the experimental method only produced higher absolute item discrimination values ten out of the twenty-five times. (Sign Test not significant). Table 5 presents item discriminations for Test 3. In twenty out of twenty-five item discriminations, the absolute value was higher for the experimental scoring procedure. Unfortunately, one cannot truly interpret these item discrimination results since the computer program employed for calculating R only prints out R^2. To arrive at R, the square root of R^2 was taken; therefore, all of the R presented in Tables 3, 4, and 5 are positive values and we did not determine if any of these values should have been negative. Since negative item discrimination values are not desirable, and since we could not discern which items, if any, should have been negative for the experimental method of

grading, the results in Tables 3, 4, and 5 should be looked at cautiously. (However, one should note that only 2 items of the 75 scored traditionally produce negative values).

Since the experimental method of grading required that the students respond twice to every test item, it was felt that this method may have produced a different testing situation which would result in different overall test scores. This was originally hypothesized by one of the authors while administering the test. He observed students verbal and non-verbal behavior indicating that they found the experimental testing procedure to be much more difficult. In the summer, 1973, to check on this possible effect, the authors randomly assigned the two different grading procedures to each of half of the two class sections of undergraduate tests and measurements. In each section, half of the students were taking the test traditionally and the other half of the students were taking it experimentally. Both tests were then graded, using the traditional grading procedures. These results are presented in Table 6.

The mean number of right answers for both procedures was approximately 18, and the standard deviation for the traditional procedure was approximately 3.4, and 3.0 for the experimental. These results indicate that the two procedures are not producing different testing situations.

The results of this study may have been unable to fully demonstrate the potential increase in effectiveness of the

experimental grading over the traditional method because some of the validity criteria (objectives and project) were lost. This loss was partially due to the students being given access to their projects which resulted in some just taking their project. A quick evaluation indicated that the projects that tended to be taken were the ones receiving the lowest test grades. This may have seriously affected our range of scores. Since our theoretical position was that the experimental method would be more sensitive in detecting partial knowledge and would therefore be better able to detect differing ability levels, then restricted ranges would severely handicap the experimental method's ability to demonstrate its effectiveness.

One should note when reading the results that shrinkage estimates were employed for the total test validity results, but they were not calculated for item validities that were reported. This should be taken into account when interpreting the results. The item validities can be found in Appendix A, and it was felt that the total test validities were of greater importance.

One should also note that the item discrimination using the multiple regression procedures were not corrected for shrinkage. This was not done because of a time factor but they theoretically should be calculated. However, one should also consider that the standard method (r) used to calculate item discrimination and item validities have not been, and generally are not corrected for shrinkage.

Another consideration, as pointed out by Uhl and Eisenberg (1970) and Newman (1973), is that there are variations between shrinkage estimate formulas. Wherry's formula, which is most commonly used, was employed for calculating shrinkage estimates for this study. One should consider using Lord's (1950) formula for a shrinkage estimate for both R and r.

In this study, an attempt was made to develop a multivariable approach for improving item validities. It seems that if such an approach is further explored one would also have to develop multivariable and multivariate[1] methods for determining reliability. If one developed a multivariate technique for improving item discrimination and item validity and still used the traditional univariable technique for calculating reliability, this would be highly inconsistent. We would like to suggest that a modification of the canonical correlation procedure may be appropriate for developing a multivariate technique for estimating reliability which would be consistent with the approach suggested in the paper for improving validity.

In conclusion, we believe that multiple regression procedures will allow one to maximally use the available existing information produced by the probabilistic responses from examinees to determine validity estimates. The traditionally-used univariable technique will only produce one weight which is calculated to maximize it's prediction. Therefore,

it is potentially much less effective than a technique that is capable of calculating a number of separate weights for maximizing prediction. In addition, working with univariable techniques may tend to fixate researchers to thinking in univariable terms, while in our estimation, multivariate and multivariable techniques are less confining and therefore are more likely to facilitate more creative and potentially more useful research. We believe multiple regression gave us the freedom which helped us conceptually derive a potentially useful method of grading and analyzing our results.

Table #2

Validity Criterion (Project Scores)

Test	Method 1 (Trad. r)	Method 2 (Exp. r)	Method 3 (Exp. R)	Method (Exp. R
1 (N=54)	.110	.37	.267	.187
2 (N=52)	.041	.204	.299	.229
3 (N=55)	.237	.198	.315	.253

Note: See Table #1 for descriptions of methods

Table #3
Item Discriminations for Test #1

Items	Traditional Scoring (r) pt. Bis	Experimental Scoring (R)	Items	Traditional Scoring (r) pt. Bis.	Experimental Scoring (R)
1	.339	.309	14	.421	.489
2	.159	.143	15	.404	.381
3	.265	.301	16	.157	.261
4	.202	.297	17	.066	.103
5	.430	.356	18	.076	.238
6	.218	.281	19	.275	.427
7	.437	.317	20	.066	.179
8	.437	.340	21	.347	.320
9	.260	.087	22	.456	.394
10	.212	.214	23	.479	.547
11	.282	.293	24	.360	.432
12	.454	.484	25	.390	.354
13	.425	.441			

Note: N=75

Table #4
Item Descriminations for Test #2

Items	Traditional Scoring (r) pt. Bis.	Experimental Scoring (R)	Items	Traditional Scoring (r) pt. Bis.	Experi Scori (R)
1	.055	.444	14	.101	.412
2	.355	.334	15	.150	.244
3	.358	.309	16	.392	.222
4	.437	.391	17	-.040	.246
5	.438	.380	18	.436	.315
6	.435	.181	19	.318	.311
7	.058	.200	20	.433	.367
8	.226	.214	21	.585	.408
9	.517	.337	22	.431	.520
10	.375	.348	23	.417	.218
11	.111	.352	24	.481	.401
12	.354	.260	25	.133	.141
13	.131	.309			

Note: N=75

Table #5
Item Discriminations for Test #3

Items	Traditional Scoring (r) pt. Bis	Experimental Scoring (R)	Items	Traditional Scoring (r) pt. Bis.	Experimental Scoring (R)
1	.185	.180	14	.206	.441
2	.148	.786	15	.0	.649
3	.188	.183	16	.285	.333
4	.279	.553	17	.294	.413
5	.172	.757	18	.315	.248
6	.402	.353	19	.379	.670
7	.229	.794	20	.431	.493
8	.370	.766	21	.069	.232
9	.523	.796	22	.162	.626
10	.604	.527	23	.370	.637
11	.054	.783	24	.323	.653
12	.478	.520	25	.306	.669
13	.155	.798			

Note: N=75

Table #6

Data from Summer Session 1, 1973
Controlling for Testing Situation Effect
for Sections 1 and 2 Combined

	Traditional Testing Situation	Experimental Testing Situation
S	3.4280	3.0220
$\bar{X}$	18.4137	18.1515
N	29.	33.

Note: No test of significance was run since the data obviously would be nonsignificant at our alpha level of .05.

A P P E N D I X A

APPENDIX A

Item Validity-Criterion: Objective

Test 3

#	Pt. Bis.	r	R	R^2	#	Pt. Bis.	r	R	R^2
1	-.1702	-.1773	.178	.0315	14	.1024	.1544	.195	.0380
2	-.1201	-.1201	.120	.0144	15	0.0	0.0		0.0
3	-.0380	.0306	.135	.0182	16	.1376	.0721	.233	.0541
4	.1847	.0854	.316	.0988	17	.1007	.0383	.201	.0403
5	.0863	.0863	.087	.0075	18	-.0208	-.0143	.027	.0007
6	-.0334	-.0806	.120	.0143	19	.1365	.1365	.136	.0186
7	.1169	.0764	.136	.0185	20	.2015	.1244	.264	.0700
8	-.0847	-.1354	.176	.0311	21	-.0156	.1287	.148	.0219
9	.1913	.2576	.363	.1314	22	.3117	.3117	.312	.0972
0	.1782	.1226	.180	.0325	23	.1278	.0277	.292	.0854
1	-.0388	-.0388	.039	.0015	24	.1058	.0226	.149	.0223
2	-.0169	.0657	.178	.0317	25	.1807	.1975	.174	.0327
3	.0072	.0072	.010	.0001					

APPENDIX A

Item Validity-Criterion Objective

Test 1

#	Pt. Bis.	r	R	R^2	#	Pt. Bis.	r	R	
1	-.0644	-.0661	.006	.0044	14	.2081	.1678	.223	.04
2	-.2544	-.2489	.256	.0657	15	.2294	.1959	.236	.95
3	-.0203	-.0203	.02	.0004	16	.9363	.2032	.360	.12
4	-.2570	.2470	.257	.0661	17	-.0503	-.0462	.050	.00
5	.1941	.2628	.313	.0980	18	.0455	.289	.07	.00
6	.1368	.2256	.315	.0991	19	.0156	.9483	.06	.00
7	.0456	.0971	.224	.0503	20	-.1191	-.1326	.134	.01
8	-.0156	-.0228	.05	.0025	21	-.0201	-.0775	.157	.02
9	.0714	.1053	.32	.0074	22	-.0610	-.0175	.110	.01
10	.0465	.0223	.035	.0073	23	-.0063	-.0395	.078	.00
11	-.0538	-.0296	.082	.0067	24	.2235	.1737	.232	.053
12	.1315	.1817	.242	.0583	25	.2514	.3249	.353	.124
13	.3629	.3324	.364	.1322					

APPENDIX A

Item Validity-Criterion: Objective

Test 2

Pt. Bis.	r	R	R^2	#	Pt. Bis.	r	R	R^2
.3314	.3314	.365	.1332	14	.1884	.0411	.151	.0227
-.1254	-.0683	.121	.0146	15	-.0993	.1890	.137	.0189
-.1502	-.144	.250	.0627	16	-.0021	-.0539	.335	.1123
.0324	.1031	.248	.0615	17	-.0394	-.0394	.074	.0054
-.0569	-.0537	.106	.0113	18	-.0638	-.1038	.303	.0917
.0587	.0836	.107	.0148	19	.0213	.0849	.210	.0441
-.0199	-.0072	.076	.0057	20	-.2729	-.2019	.105	.0110
-.0246	.0761	.187	.0350	21	-.0747	.0130	.161	.0260
.1696	.1233	.052	.0027	22	.1820	.0830	.166	.0277
-.2151	-.1690	.112	.0126	23	.2537	.2010	.391	.1527
.0089	-.1639	.204	.0411	24	-.0605	-.0904	.047	.0022
-.1261	-.1525	.135	.0183	25	-.1850	-.1403	.201	.0404
.0090	-.1351	.273	.0743					

APPENDIX A

Item Validity-Criterion: Projects

Test 1

#	Pt. Bis.	r	R	R^2	#	Pt. Bis.	r	R
1	.1947	.1115	.310	.0963	14	.2463	.2463	.246
2	-.0677	-.0064	.201	.0405	15	.0189	.0398	.075
3	.1104	.1104	.110	.0122	16	.2119	.2342	.234
4	-.1374	-.1765	.227	.0515	17	-.2748	-.2415	.281
5	.1804	.1804	.181	.0326	18	.1383	.1223	.140
6	.0719	.0719	.072	.0052	19	-.1033	.0063	.179
7	.3221	.3222	.326	.1062	20	.1080	.1495	.182
8	.0189	.0528	.140	.0196	21	.0947	.0939	.096
9	.1336	.0928	.033	.0011	22	.0509	.0860	.111
10	-.2101	-.1630	.232	.0538	23	-.1100	-.0822	.105
11	-.0599	-.0566	.060	.0036	24	.0220	.0666	.110
12	.0036	.0114	.0002	.0006	25	-.0111	-.0017	.003
13	.0899	.0862	.090	.0081				

APPENDIX A

Item Validity-Criterion: Projects

Test 2

#	Pt. Bis.	r	R	R^2	#	Pt. Bis.	r	R	R^2
1	.2512	.2361	.318	.1001	14	.1201	.0884	.098	.0096
2	-.0341	.0635	.296	.0874	15	.1423	.1923	.179	.0321
3	.1357	.1232	.094	.0088	16	.0275	.0275	.0027	.0007
4	.2010	.2119	.224	.0501	17	-.0038	-.0038	.047	.0022
5	.1623	.1524	.171	.0291	18	-.2535	-.1848	.189	.0357
6	.2450	.2555	.242	.0586	19	-.1313	-.0570	.179	.0321
7	.1423	.1423	.179	.0321	20	-.1710	-.2092	.164	.0268
8	-.0340	.0239	.233	.0540	21	-.1167	-.0251	.284	.0808
9	.3040	.2774	.180	.0325	22	.1175	.0612	.098	.0096
0	-.0965	-.1528	.172	.0297	23	.1486	.1683	.127	.0162
1	-.2525	-.3512	.230	.0527	24	-.0747	-.1095	.106	.0112
2	-.0596	-.1635	.193	.0372	25	-.0384	-.2323	.217	.0470
3	.1222	.0075	.321	.1031					

APPENDIX A

Item Validity-Criterion: Projects

Test 3

#	Pt. Bis.	r	R	R^2	#	Pt. Bis.	r	R	R^2
1	.1656	.353	.201	.0405	14	.3654	.3678	.373	.13
2	-.1187	-.1416	.148	.0219	15	0	0		0.0
3	.0166	0	.033	.0011	16	.2160	.1397	.307	.09
4	.0431	-.0325	.211	.0445	17	.2274	.2561	.260	.06
5	.1017	.1017	.101	.0103	18	.1929	.0791	.335	.11
6	-.0371	-.0441	.045	.0020	19	.1744	.1744	.174	.03
7	-.1783	-.1009	.232	.0538	20	.1944	.2171	.219	.047
8	-.0967	-.1169	.148	.0222	21	-.0455	.0651	.065	.004
9	.0352	.0946	.143	.0205	22	.1744	.1744	.174	.030
10	.2474	.2095	.250	.0624	23	.1379	.0052	.186	.034
11	.1744	.1744	.174	.0304	24	-.0755	-.0290	.066	.004
12	.1258	.0858	.178	.0315	25	.0422	-.0020	.108	.011
13	-.0455	-.0455	.213	.0455					

Reference

Archer, N. S., "A Comparison of the Conventional and Two Modified Procedures for Responding to Multiple-Choice Items with Respect to Test Reliability, Validity, and Item Characterists." Unpublished Doctorial Dissertation. Syracuse University, 1962.

Coombs, Milholland, Womer, "The Assessment of Partial Knowledge." Educational and Psychological Measures, 1965, 16, 13-37.

Ebel, "Confidence Weighing and Test Reliability." Journal of Educational Measurement, 1965, 2, 49-57.

Echternacht, B., "The Use of Confidence Testing in Objective Tests." Review of Educational Research, 1972, 42, 2, 217-236.

Hambleton, Roberts, Traub, "A Comparison of the Reliability and Validity of Two Methods for Assessing Partial Knowledge on a Multiple-Choice Test." Journal of Educational Measurement. 1970, 7, 75-82.

Jocobs, Stanly S., "Correlation of Unwarranted Confidence in Responses to Objective Items." Journal of Educational Measurement, Vol. 8, sp 1971.

Kelly, Beggs, McNeil, Eichelberger and Lyon, Research Design in the Behavioral Sciences: Multiple Regression Approach, Southern Illinois University Press, 1969.

Koehler, Roger O., "A Comparison of the Validities of Conventional Choice Tests and Various Confidence Marking Procedures." Journal of Educational Measurement, Vol. 8, No. 4, Winter, 1971.

Lord, Novick, Statistical Theories of Mental Test Scores. Addison-Wesley, 1968.

Newman, Isadore, "Variations Between Shrinkage Estimation Formulas and the Appropriateness of Their Interpretation." Multiple Linear Regression Viewpoints. Vol. 4, 2, 45-48, 1973.

Uhl, N. and Eisenbert, T., "Predicting Shrinkage in the Multiple Correlation Coefficient." Education and Psychological Measurement, 30, 487-489, 1970.