Florida Journal of Educational Research Fall 1995, Vol. 35, No. 1, Pp. 54-62.

Item Analysis of Criterion-Referenced Tests

L. R. Gay and Jo D. Gallagher
Florida International University

ABSTRACT. We propose that teachers should not calculate reliability but rather should use item analysis data as the major determinant of the adequacy of classroom measures. Using the results of two actual classroom tests--one elementary-level mastery test and one secondary-level nonmastery test--we demonstrate an approach to item analysis for CRTs that involves the calculation of item difficulty for all items *and* discriminating power for those items with unacceptable *P* values. For each test, the effects on initial reliability estimates and item analysis indexes of systematically removing low-scoring students' results (one at a time) are presented. We conclude that this modified item analysis strategy for classroom CRTs is both efficient and useful.

Discussion in most measurement texts either states or implies that there are two kinds of tests, norm-referenced tests (NRTs) and criterion-referenced mastery tests; at the very least the suggestion is that *most* criterion-referenced tests (CRTs) are mastery tests. Even sources that acknowledge the existence of nonmastery CRTs present related item analysis information in a mastery context. A statement is usually made to the effect that since the purpose of a CRT is to describe what students have learned, not to discriminate among them, and since we are not interested in promoting variability, indexes of discriminating power are of little or no value.

Most classroom tests, however, do have some degree of variability, intended or not. It is rare, even for a mastery CRT, to have *all* items with "acceptable" *P* values. We propose that for items with unacceptably low *P* values (based on teacher judgment), indexes of discriminating power can provide valuable information for revision of items or instruction. An item with a *P* value of 50, for example, which discriminates well between high and low achievers (e.g., $D = .60$), may indicate simply that a more difficult concept was measured. An item with a *P* value of 50 and a *D* value of .00 or $-.20$, on the other hand, suggests that there is a problem somewhere.

Reliability is just as important for criterion-referenced classroom tests as for other tests. There is currently, however, no satisfactory approach for calculating it. It is usually suggested that since CRTs exhibit little or no variability, traditional methods for computing

reliability should either be avoided for such tests or applied with great caution and interpreted differently. This premise is not necessarily true; some classroom CRTs have fairly large standard deviations (SD) and high reliability coefficients. It is true, however, more often than not. Classroom test score distributions tend to be negatively skewed; even nonmastery CRTs generally have less variability than typical NRTs. As variability decreases, so does the related reliability coefficient.

Thus, if most students achieve most intended outcomes, it is very possible to have a highly reliable test which yields a coefficient near .00. Also, since most classroom test distributions are negatively skewed and based on a relatively small number of test takers, removal of results for even one student can dramatically affect the variability and hence the computed reliability coefficient. This can be empirically demonstrated with actual classroom test data.

Suggested alternatives to traditional reliability coefficients typically involve some type of analysis of test-retest differences, usually consisting of mastery-nonmastery decisions or analysis of size of score discrepancies for nonmastery tests. The method suggested by Swaminathan, Hambleton and Algina (as cited in Berk, 1980), for example, although computationally reasonable, requires two testings and produces errors of estimation that are somewhat large for sample sizes found in classrooms. Others, such as those suggested by Hyunh, Subkoviak, and Marshall-Haertel, respectively (as cited in Berk, 1980), involve only one test administration, but are computationally wearisome and yield biased estimates for short tests (Subkoviak, 1980). For teacher-made or -selected classroom tests, Berk (1984) recommends the p_o index: "the proportion of individuals consistently classified as masters and nonmasters of an objective based on a threshold or cut-off score" (p.235). He goes on to say that Hambleton and Novick's (as cited in Berk, 1984) test-retest method for estimating p_o is the simplest to understand and compute. But it also has the same drawbacks as the Swaminathan et al method. Such alternatives, it appears, are not very practical for the average classroom teacher for whom it is not typically feasible to administer tests twice (either identical or parallel) or to have the computer capability to facilitate complex computation of coefficients. On the other hand, neither does it seem satisfactory to tell teachers, as is often done, that teacher-made tests usually produce traditional reliability coefficients in the .50 range and that such coefficients are acceptable, given the decisions that will be made based on the results. This approach sends a confusing, mixed message concerning what constitutes an acceptable level of reliability.

Given the above problems with estimating the reliability of classroom tests using various methods developed for CRTs, and the known relationship between traditional (NRT) reliability coefficients and item analysis results, we propose that perhaps the best solution is to discourage teachers from computing reliability coefficients and to instead promote appropriate item analysis—an analysis that involves computing "traditional" P and D values, but interpreting them in a way more useful to the classroom teacher. While some may argue that teachers need not perform any calculations, we believe that they should have at least one empirical tool for evaluating the quality of their tests and related

instruction. For a NRT we can verify that if all or most items have average *P* values and acceptable *D* values, reliability coefficients will be high. Similarly, if a CRT is valid and contains mostly well-constructed items, its reliability will be high, although there may be no satisfactory way to verify it. Further, it can be demonstrated with actual classroom test data that item analysis conclusions are more stable than reliability estimates; removing results for the lowest-scoring student, for example, may dramatically affect the latter, but not the former.

Method

The proposed approach to item analysis of criterion-referenced classroom measures involves the following steps:

1. Prepare a student-by-item matrix which indicates only incorrect responses with an X (or some other symbol of choice, such as a 0). As usual, list students in order of total score, highest to lowest. Label item clusters which relate to the same objective (see Figure 1 for an example).

2. Calculate the item achievement rate (*P* value) for each item; include results for all students in the calculations.

3. For each item, make a judgment concerning the acceptability of the *P* value. The criterion for acceptability of the *P* value will vary depending upon such factors as whether the item is intended to measure a mastery objective and the difficulty of the concept measured. Thus, for example, for certain items, $P = 70$ might be considered acceptable.

4. For items with unacceptable *P* values (and *only* for items with unacceptable *P* values), calculate discriminating power (*D*), based on the results for high and low scoring subgroups (e.g., top third, bottom third).

5. Logically analyze the results.

While there is some disagreement as to how many students' results should be included, for CRTs we advocate calculating *P* values based on all students' results for the following reasons:

1. With classroom tests, teachers deal with a relatively small number (15-40) of test takers and they want to know how the class performed as a whole. It thus makes more sense to base *P* values on total group performance.

2. The small number of students available for statistical purposes affects the stability of the item indices. Using the results of the total group provides as much stability as is possible.

Additionally, some may feel that *D* values should be calculated for all items regardless of their *P* values. As a practical matter, however, the information gained from calculating *D* values for items with *P* values of 90, for example, is of questionable value for a CRT.

We selected two actual teacher-made tests to demonstrate the above procedure: an elementary-level mastery math test involving addition math facts with sums to ten (see Figure 1), and a secondary-level nonmastery calculus test (see Figure 2). Data were obtained through normal use of these tests in the respective teachers' classrooms.

Results

Results were analyzed for all students in each class (*n*), for the total group minus the lowest scorer (*n* -1), and for the total group minus the two lowest scorers (*n* - 2). This was done to examine the stability of various values since, for any given test, one or more students may be absent or inappropriately placed in the classroom (e.g., should be in a special class).

For the mastery test, the performance level of the class was fairly high (M = 17.83, K = 20) (see Table 1) and only two items had *P* values (72) which warranted calculations of *D* values (see Table 2). In both cases, the items discriminated well. Visual analysis of the items revealed that they were structurally identical to other items which also dealt with "missing addends" (e.g., __ + 5 = 8) and for which *P* values ranged from 83 to 100. Interestingly, both items involved 4 as the missing addend. The KR-21 and KR-21′[1] estimates were .90 and .88, respectively. Progressively removing the two lowest-scoring students, not unexpectedly, resulted in an increase in the mean and a decrease in the SD; the effect on the reliability estimates was, however, dramatic. The KR-21, for example, dropped from .90 to .05.

For the nonmastery test, the overall performance level of the class was not as high as for the mastery test (M = 22.31, K = 27), and the SD was lower (see Table 1). For this test, eight items had *P* values which warranted calculation of *D* values (see Table 2). All of the items discriminated well (D = .40 to .80) except item 21 (D = −.20). Interestingly, item 21 was structurally identical to item 22 for which P = 100. The teacher who constructed the test believed that it was the solid dot on the related graph which had confused both high-scoring and low-scoring students. In other words, the item was fine but the graph was confusing. As would be expected, the initial reliability estimates were not as high (KR-21 = .59, KR-21′ = .65). As with the mastery test, progressively removing the two lowest-scoring students resulted in an increase in the mean, a decrease in the SD, and significant decreases in the reliability estimates.

Performance Outcome	1 1/3				2 3/4				3 5/6						4 5/6						
Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Total Score
Student 1																					20
2																					20
3																					20
4																					20
5																					20
6																					20
7																					20
8																					19
9																					19
10																					19
11																					18
12																					18
13																					18
14																					18
15																					18
16																					17
17																					12
18																					5
P	100	100	94	100	94	89	78	94	83	100	72	83	72	89	83	94	89	89	94	83	
D											.67		.50								
OAR		1	0	0		9	4				8	3					8	3			

Figure 1. Student-by-item matrix for elementary-level mastery math test.

	ITEMS																											
S	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	TOTAL SCORE
1							X																					26
2				X																								26
3									X																			26
4																		X		X								25
5																												26
6		X																			X				X			24
7								X										X		X	X							23
8			X			X	X		X												X							22
9			X			X				X			X		X													22
10							X	X							X						X						X	22
11		X			X						X															X	X	22
12						X	X	X											X	X							X	21
13			X		X	X	X		X																X		X	20
14			X		X			X		X					X				X	X								20
15		X	X	X						X	X			X	X			X		X								18
16		X	X			X	X	X	X		X				X			X	X	X							X	15
P	100	75	63	88	81	69	63	69	75	81	81	100	94	94	69	100	100	75	81	63	69	100	100	100	88	94	63	
D			.80			.60	.40	.60							.60					.60	-.20						.40	

Figure 2. Student-by-item matrix for secondary-level nonmastery calculus test.

Table 1
Descriptive Statistics and Reliability Estimates for an Elementary-Level Mastery Test and a Secondary-Level Nonmastery Test

Students	*M*	SD	KR-21	KR-21'
Mastery Test: Addition Facts with Sums to 10[a]				
1 - 18	17.83	3.62	.90	.88
1 - 17	18.59	1.91	.68	.71
1 - 16	19.00	1.00	.05	.24
Nonmastery Test: Calculus[b]				
1 - 16	22.31	2.99	.59	.65
1 - 15	22.80	2.40	.40	.51
1 - 14	23.14	2.10	.26	.40

Note. For both tests students were arranged by order of score, highest to lowest.
[a]Maximum Score = 20. [b]Maximum Score = 27.

With respect to item analysis for the mastery test, the corresponding effects of progressively removing the two lowest scoring students were increased *P* values and decreased *D* values (see Table 2). Even when the *P* values reached 81, the corresponding *D* values remained positive. The corresponding effects on the item analysis data for the non-mastery test were, in general, minimal. While in some cases *P* values increased and *D* values decreased, results were very stable. The items that discriminated continued to do so and the item that discriminated negatively (item 21) continued to either discriminate negatively or not discriminate at all (see Table 2).

Conclusions

The data confirm that for both mastery and nonmastery classroom CRTs, a modified approach to item analysis is efficient and produces conclusions which are more stable than those resulting from reliability estimates. For the mastery test, only two of 20 items suggested calculation of *D* values, and for the nonmastery test, eight of 27. In general, removing data for the two lowest-scoring students tended to 1) increase *M* and *P* values and decrease SD and *D* values (as would be anticipated), and 2) greatly decrease reliability estimates. Assessment of item adequacy, however, tended to stay the same. Shifts were somewhat more dramatic for the mastery test. The teacher of the group involved indicated that the two lowest-scoring students had been referred for transfers to special classes. In summary, item accuracy conclusions and, therefore, the decisions teachers make based on them remain relatively stable given the presence or absence of a few students. This is in contrast to reliability estimates and the decisions based on them. Thus we recommend that teachers should not calculate any type of reliability estimate for a CRT and, instead, should perform a modified item analysis as we have described here.

Table 2

Item Analysis Indices for Low P Value Items on an Elementary-Level Mastery Test and a Secondary-Level Nonmastery Test

Student	Mastery Test: Addition Facts with Sums to 10[a] Item			
	11		13	
	P	*D*	*P*	*D*
1 - 18	72	.67	72	.50
1 - 17	76	.50	76	.33
1 - 16	81	.33	81	.17

	Nonmastery Test: Calculus[b] Item													
	3		6		7		8		15		20		21		27	
	P	*D*	*P*	*D*	*P*	*D*	*P*	*D*	*P*	*D*	*P*	*D*	*P*	*D*	*P*	*D*
1 - 16	63	.80	69	.60	63	.40	69	.60	68	.60	63	.60	69	-.20	63	.40
1 - 15	67	.60	73	.40	67	.20	73	.40	73	.40	67	.40	67	-.20	67	.40
1 - 14	71	.40	71	.40	64	.40	71	.60	79	.40	71	.20	64	.00	64	.60

Note. All *P* values are based on the achievement of all students. For the matery test, *D* values are based on the achievement of the top six and bottom six students. For the nonmastery test, *D* values are based on the top and bottom five students.
[a]Maximum Score = 20. [b]Maximum Score = 27.

We recognize that teachers do not routinely perform item analysis. Those who have applied the modified approach we suggest, however, report that it does not take much time and that it gives them valuable input, not only for improving their tests, but also for enhancing their teaching. Further, they claim to enjoy the process and sharing results with students. Some teachers have discovered that the scoring machines they use from time to time provide the item statistics and that the microcomputers in their schools and homes are good tools for preparing student-by-item matrices. Lastly, they report that while they may not always do all the calculations, they often apply the procedure informally, a strategy we believe should be encouraged.

Note

[1]As the KR-21 estimate is generally very conservative, i.e., low, the KR-21′ formula was developed to get a closer approximation of the KR-20 value while maintaining the ease of application of the KR-21. The coefficients in Table 1 essentially bear this out (with the exception of one which was high even by KR-21 standards—that for the mastery test for all 18 students).

References

Berk, R. A. (Ed.). (1984). *A guide to criterion-referenced test construction.* Baltimore, MD: Johns Hopkins University Press.

Berk, R. A. (Ed.). (1980). *Criterion-referenced measurement: The state of the art.* Baltimore, MD: Johns Hopkins University Press.

Subkoviak, M. J. (1980). Decision consistency approaches. In R. A. Berk, (Ed.), *Criterion-referenced measurement: The state of the art* (pp. 129-185). Baltimore, MD: Johns Hopkins University Press.

Jo D. Gallagher, DM 291, Florida International University, Tamiami Trail, Miami, Florida 33199.