REVISED GUIDELINES FOR CONDUCTING ITEM ANALYSES OF CLASSROOM TESTS

Dale Shaw
University of Northern Colorado

Suzanne Young
University of Wyoming

Twenty-seven multiple-choice, instructor-created classroom tests from a wide variety of college subjects were submitted to an iterative form of item analysis. Without exception, the tests contained large numbers of items that contributed little or nothing to the measurement of the test taker's knowledge. On average, the best one-third of items (those with the highest item-to-total score correlation coefficients) from each test, discriminated among test takers as well as the whole test did. The paper presents revised guidelines for conducting item analyses to improve classroom tests that call for cutting deeper than current published recommendations when identifying items to be replaced or rewritten.

IMPROVING THE RELIABILITY OF CLASSROOM TESTS

Teachers who create objective tests for use in their own classrooms may draw upon a substantial collection of well known test development, analysis, and revision practices that, if used thoughtfully, will produce tests that have acceptable reliability and validity. Tests and measurements textbooks such as Sax (1997) and Thorndike (1997) suggest that tests should consist of items that meet certain standards of clarity and that cover the content domain appropriately. Hopkins (1998) adds that, once a test has been carefully designed, "[item analysis] can be a valuable activity that can enhance the test's reliability and validity" (p. 255). Upon administration, if the test is to be used again, the items should be submitted to item analysis and "weak" items revised or replaced. The resulting version of the test should be an improved test.

This paper deals with the item analysis phase of classroom test development. The focus of this work is to bring under scrutiny what seem to be usual test construction practices of classroom teachers with the intent to see if these practices can be modified or refined so as to produce better tests. Sax (1997) states, "Devote more time to writing and editing items than to running the item analysis. Little is gained by expending all of one's time, energy, and resources for an item analysis that leaves no time to improve items or help students." There is much truth to this. In test construction, there is no substitute for well-crafted items. However, an item analysis will assist us greatly in identifying items that should be revised as well as items that should be left as they are. Indeed, if the 27 tests that we analyzed in this paper are typical of classroom tests in general, the item analysis procedures and practices that we are recommending to use will enhance the reliability of classroom tests substantially.

When we, as classroom teachers, set out to create an objective test, we are able to use well-known test development practices to create a test that we are reasonably confident is a valid test of the content domain. We use pre-written items when they are available, modify these pre-written items to better sample the content as we have covered it in the classroom, write carefully crafted items when pre-written ones are unavailable, and select or create items that cover the domain of content appropriately. We do not expect our test to be perfect, but we do sufficient work to allow ourselves to believe that the test's content validity is reasonably high. We deal with

security issues so that we can conduct an item analysis, make revisions to the test as warranted, and use it again. We administer the test and submit it to item analysis even if the class is not large in size.

A typical item analysis begins with a printout from any of a number of standard computer programs that scores the tests and provides item information. Coefficient alpha is computed for the whole test and item-to-total score correlation coefficients are computed for each item. Presented with each item-to-total score correlation is a revised alpha coefficient that is computed for the test with that item deleted. Our purpose in this study was to seek ways in which we might better use this sort of item analysis information to improve classroom tests. The issues we addressed include both practical and conceptual matters.

With regard to practical matters, the typical item analysis informs us only about the resulting test in which a single item is deleted or removed. In this study we asked about the properties of the resulting test when a group of items is removed. What could an analysis tell us, for example, if the worst two items or worst three items were deleted or replaced? Or the worst six items? How about the worst 1/3 or 1/2 of items? Would this type of additional item analysis information help us understand better how deeply to cut in identifying items to be revised or replaced? Would this information help us to determine if there is a realistic item-to-total score correlation cut-off value that is appropriate for most or all classroom test construction situations?

The conceptual matters are developed later in this paper. Essentially these matters relate to each item's ability to discriminate among test takers, each item's ability in the presence of the other items to add to the test's ability to discriminate among test takers, and how we may view these sorts of discriminations as captured within item-to-total score correlation coefficients, coefficient alpha for the test, and item and test validities.

PROCEDURES

Item analyses were conducted on 27 multiple-choice, instructor-created classroom tests that had been developed and administered in college courses in a wide variety of subjects and disciplines by 21 different instructors. Instructors were asked for data from objective tests that they had created as midterm or unit tests in which speed was not a factor. They volunteered the data from these tests with all identifying student information deleted, and with full knowledge of the purpose of the project. Some of the instructors had used test bank items exclusively, some had written their own items, and some had used a combination of these two item sources. Most of the instructors who had used pre-written items from test banks said that they had had to rewrite or refine many of the items they used. All but three of the instructors said they were well acquainted with objective test development procedures. All instructors indicated that they thought they had done a reasonably careful job in creating a valid test. All tests were first-time creations and had not been submitted to item analysis nor revised previously. Class sizes ranged from a low of 18 students to a high of 46 students.

A computer program was developed to perform an iterative item analysis that permitted us to analyze those items that remained in a test after discarding the least discriminating item, then those items that remained after discarding the next least discriminating item, then the next least discriminating item, and so on. Coefficient alpha was computed for the test, item-to-total score correlation coefficients were computed for all items with the total score, and the least discriminating item based on these correlation coefficients was discarded. Coefficient alpha and item-to-total score correlation coefficients were recomputed for the remaining items and the next least discriminating item was identified and discarded. This iterative process was continued until only two items remained. Program output included tables and graphs presenting the recalculated alpha coefficient and the item-to-total score correlation coefficient for the item discarded at each step. Tables and graphs were prepared to summarize the results of the 27 tests that we analyzed.

RESULTS

The results for the first test that we analyzed are presented in Table 1. These results are typical of the results we obtained for the other 26 tests. Indeed, the results that we obtained for all 27 of the tests were so similar that we could have used any one of them here to illustrate our processes and analyses. Figure 1 depicts coefficient alpha plotted against the number of items remaining in the test as items were discarded one by one.

The curve that was generated was essentially parabolic in nature, flattened somewhat on top, and opening downward. Items that correlated negatively with total score were the first ones discarded in the iterative item analysis. As these items were discarded one by one, coefficient alpha increased slowly. Coefficient alpha was 0.735 for the entire test of 50 items. After discarding the seven items that correlated negatively with total score, coefficient alpha had increased to 0.790. Items correlating near zero and those with low positive correlations with total score dropped out next; all the while, coefficient alpha was increasing slightly along the ever-flattening top of the curve. With 22 items removed (almost half of the items), coefficient alpha reached its highest value, peaking at 0.814. To this point, item-to-total score correlation coefficients had ranged from a low of -0.230 to a high of +0.203 for the discarded items. The next portion of the curve contained perhaps the most interesting information. As items continued to be dropped, coefficient alpha began to decrease as

Table 1
Coefficient Alpha and the Correlations of Items Deleted with Total Score

Number of Items Deleted	Coefficient Alpha	Correlation of Item with Total Score	Number of Items Deleted	Coefficient Alpha	Correlation of Item with Total Score
0	0.735	-0.230	21	0.814	0.201
1	0.755	-0.192	22	0.814	0.203
2	0.759	-0.142	23	0.811	0.237
3	0.772	-0.077	24	0.808	0.238
4	0.779	-0.019	25	0.805	0.246
5	0.786	-0.004	26	0.802	0.260
6	0.788	-0.004	27	0.798	0.265
7	0.790	0.000	28	0.796	0.266
8	0.790	0.000	29	0.791	0.269
9	0.791	0.000	30	0.790	0.282
10	0.791	0.000	31	0.777	0.283
11	0.792	0.000	32	0.770	0.289
12	0.792	0.043	33	0.767	0.292
13	0.798	0.056	34	0.759	0.301
14	0.800	0.078	35	0.753	0.302
15	0.800	0.107	36	0.731	0.302
16	0.808	0.108	37	0.722	0.315
17	0.808	0.125	38	0.714	0.315
18	0.810	0.127	39	0.708	0.334
19	0.811	0.150	40	0.671	0.335
20	0.813	0.194	41	0.654	0.351

expected. When alpha had decreased to the point approximately where it had started (alpha = 0.735 for the 50-item test), only 14 items remained in the test. With regard to reliability and discrimination among subjects, these 14 items were essentially doing the work of the original 50 items! The item-to-total score correlation for the 36th item removed was 0.302. As Figure 1 illustrates, deleting the remaining 14 items one-by-one caused alpha to decrease rapidly.

A summary of the results for all 27 tests is presented in Table 2. Initial coefficient alphas for these tests ranged from 0.605 to 0.875. Except for one test, 1/3 to 1/2 of the items in all of the tests could be deleted with

Figure 1. Relationship of Number of Items Deleted and Coefficient Alpha

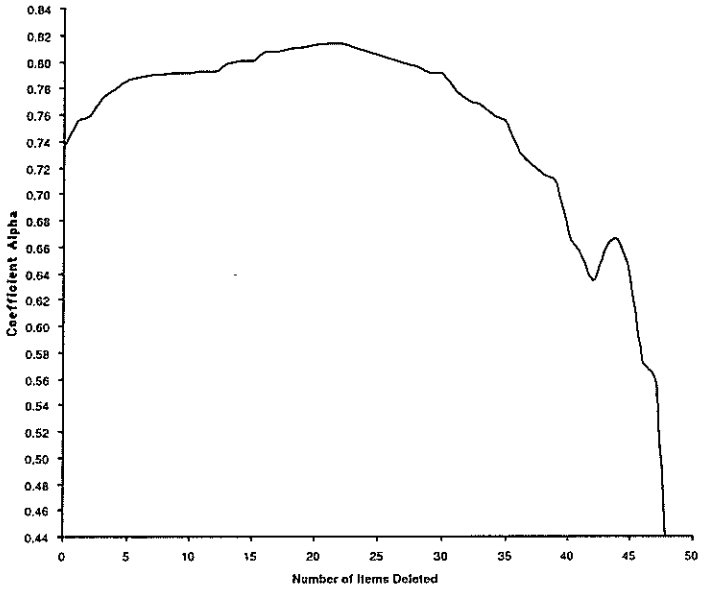
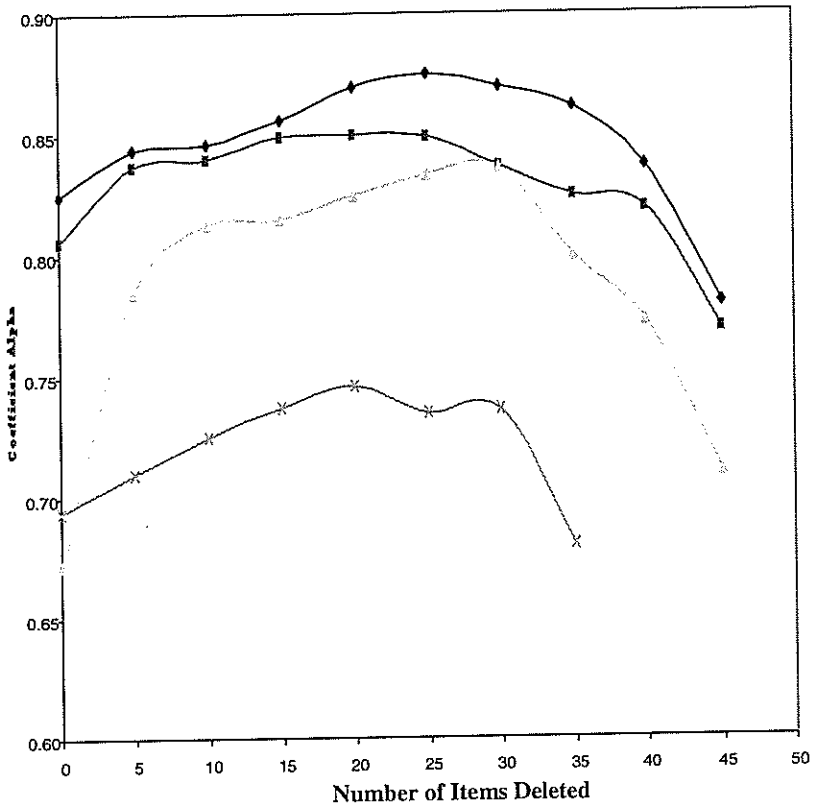


Table 2
Coefficient Alpha at Starting Point, Leveling Point, Highest Point, and Equal to Starting Point

Test	α at starting point	No. of items in test	α at leveling point	No. of items in test	α at highest point	No. of items in test	α = to starting point	No. of items in test
Test #1	0.735	50	0.808	34	0.814	28	0.735	14
Test #2	0.816	50	0.827	42	0.836	30	0.816	20
Test #3	0.806	50	0.847	38	0.851	31	0.807	16
Test #4	0.672	50	0.827	30	0.844	23	0.618	4
Test #5	0.825	50	0.865	31	0.875	24	0.817	7
Test #6	0.795	50	0.865	32	0.874	27	0.790	8
Test #7	0.694	40	0.740	25	0.748	20	0.704	16
Test #8	0.797	40	0.845	24	0.874	16	0.778	5
Test #9	0.822	40	0.866	26	0.879	16	0.821	7
Test #10	0.814	40	0.836	31	0.837	25	0.814	11
Test #11	0.878	50	0.898	37	0.906	20	0.879	10
Test #12	0.759	40	0.842	27	0.889	14	0.732	9
Test #13	0.723	40	0.796	29	0.828	20	0.744	11
Test #14	0.875	40	0.895	30	0.919	10	0.882	7
Test #15	0.775	50	0.903	24	0.985	8	0.802	4
Test #16	0.791	45	0.888	23	0.970	8	0.842	5
Test #17	0.718	50	0.778	41	0.781	28	0.731	10
Test #18	0.724	50	0.760	37	0.770	29	0.732	11
Test #19	0.794	40	0.836	24	0.847	16	0.791	7
Test #20	0.674	45	0.783	24	0.852	17	0.690	9
Test #21	0.711	50	0.806	29	0.877	20	0.744	11
Test #22	0.839	50	0.911	36	0.950	26	0.828	18
Test #23	0.847	40	0.892	24	0.927	16	0.848	10
Test #24	0.803	40	0.873	28	0.898	25	0.800	18
Test #25	0.795	50	0.829	36	0.868	27	0.811	15
Test #26	0.847	50	0.914	33	0.946	28	0.851	18
Test #27	0.605	40	0.795	26	0.852	19	0.613	8

concomitant increases in their alpha coefficients. Alpha invariably peaked with approximately 1/2 of the items removed. Most remarkable was the finding that about 1/3 of the items (those items with the highest item-to-total score correlations) in all 27 tests could alone produce an alpha value equivalent to that of the test of original length. The degree to which the results varied from test to test and the subtleties of these variations are illustrated in Figure 2. Four other tests are plotted in Figure 2; three 50 -item tests and one 40-item test are depicted. It was of interest to note that tests with high starting alpha coefficients exhibited the same characteristics in their graphs as tests with low starting alpha coefficients.

Figure 2. Relationship of Number of Items Deleted and Coefficient Alpha for Four Tests



INTERPRETATION AND DISCUSSION

We begin this section with four practical recommendations for constructors of classroom tests. Three of these recommendations are consistent with typical textbook suggestions with some minor refinements and greater specificity, and one goes beyond these typical textbook guidelines. Based on our analysis of the 27 classroom tests described in this study, these recommendations constitute good practice when conducting an item analysis of an objective, classroom test.

Recommendations

Delete, replace, or rewrite all items with negative item-to-total score correlation coefficients. No exceptions.

This is consistent with advice in all measurement textbooks. There is nothing to recommend retaining such items. In Figure 1, the first seven items to be deleted have negative item-to-total score correlation coefficients. These seven items discriminate among test takers somewhat, but they do it in an opposite way compared to the bulk of the items. Simply deleting these items causes coefficient alpha to increase appreciably from 0.735 to 0.790. The same may be seen for the other 26 tests we analyzed.

No textbook author recommends retaining an item having a negative item-to-total score correlation. Even if the item pertains to a point or idea in the content domain that one believes is essential or of critical importance, it is not a good item. We belabor the obvious here because it serves as a clear example of one of the conceptual issues. An item that correlates negatively with the total test score does not itself discriminate among test takers properly, it does not contribute to the test's ability to discriminate among test takers well, and, most importantly, it creates error in the measurement of subjects that is reflected in a reduction in the alpha coefficient. Such an item is not a valid item for this test. It does not add to the test's ability to measure a test-taker's knowledge of the domain of content.

Merely deleting the items with negative item-to-total score correlation coefficients will enhance the reliability of a test. In Figure 1, the

removal of the seven items having negative item-to-total score correlations (14% of the items in the whole test) increases coefficient alpha and causes the test to be improved. It will be a more valid test of the content domain with the items removed than with them retained! Forty-three good items will provide us with a more valid test of the content domain than the same forty-three items and seven negatively discriminating items adding error into our measurement. Admittedly, the test is not as valid as it could have been had the seven items sampled their bits of the content domain well. The point is simply that deleting the items without replacing them will improve the test. Replacing or revising these items will improve the test even more. Obviously we should replace negatively discriminating items, not just delete them.

Replace or rewrite all items with zero correlation coefficients with total test scores.

As illustrated in Figure 1 and Table 1, simply deleting these items causes alpha to increase slightly. We found this to be true for the other 26 tests as well. These items should be rewritten or replaced, however. The opportunity to replace 5 items that are providing no discrimination whatsoever in the scores of test takers, can lead to substantial improvements in the test's validity and reliability. In an objective test, these items are contributing nothing to the measurement of the construct and should be replaced.

Replace or rewrite all items with low positive item-to-total score correlations unless there are very few of them and there are no items like those described in 1) or 2) above.

As illustrated in Figure 1, items with low positive item-to-total score correlation coefficients between 0.00 and 0.20 can be discarded from the test and coefficient alpha will rise. We found this to be true for the other 26 tests as well. Deleting these items will result in a test that is slightly more reliable than a test in which they are retained. Replacing them with better items will, of course, improve the test.

For the 27 classroom tests we analyzed, coefficient alpha peaked on average after half of the items had been discarded (see Table 2). The concomitant item-to-total score correlation coefficients for the items at or near the middle mark in the iterative process were invariably very close to 0.20. Writing this recommendation in terms of a "peaking alpha" is appropriate only for persons doing an iterative item analysis of the sort described in this paper. With a conventional item analysis output, an item-to-total score correlation of 0.20 would appear to be a reasonable cut point that generally corresponds to the point where alpha peaks. Retaining the items that produced the least amount of error collectively (to the point where alpha peaks), and replacing or revising all other items having item-to-total score correlations of 0.20 or less, would appear to be excellent practice. Using a cut-point of 0.20 for items to be retained in a revised test, test constructors will produce more reliable tests that consist of greater numbers of valid items. However, we can do even better (although probably not much better), by using the recommended cut-point in 4. below.

If the worst items in a test have low positive item-to-total score correlations and there are only a few of them, then the curve one would obtain similar to Figure 1 would appear to be monotonically decreasing. It would start out on the left as essentially flat and then decline continuously as items were discarded. Discarding any of these items would do little to enhance the reliability of the test. Revising them or replacing them with items that might produce improved discrimination may prove to be fruitless. A test maker has probably achieved the best test possible without making further revisions.

Consider using 0.30 as the item-to-total score correlation cut-off point for identifying items that may merit revision.

Coupled with 3), this suggests that one considers items for possible revision with item-to-total score correlation's falling in the range of 0.20 to 0.30. These are fairly good to quite good items. They could stand as written. Improvement, if it is possible, tends to keep the

parabolic curve from bending over too rapidly, and allows the curve to rise to a higher plateau and a higher coefficient alpha value for the whole test.

DISCUSSION

An important practical matter that warrants discussion is the problem created when objective tests that are used for classroom purposes are improved to the point that all or most items discriminate well among the test takers. Such tests, although highly reliable, become quite challenging for the test takers. Focusing on improving item discrimination when revising tests tends to lead to tests that consist predominately of difficult items. This is not to say that we should back away from item improvement; it may speak more to how grading practices may have to be modified to accommodate tests being designed to create greater variance in scores.

With regard to a related matter, the items that remain in a test the longest under this type of iterative item analysis, are those items that have the highest item-to-total score correlations, have the highest discrimination powers among test takers, and tend to be the items that students report are the most challenging items. The last two columns of Table 2 contain information about the number of items that were required to produce the coefficient alpha value that was equivalent to the alpha value for the entire test. For the test presented in Figure 1, the best 16 items produced an alpha of 0.730. Adding back in the remaining 34 items caused alpha first to increase to 0.814 as items with item-to-total score correlation's between 0.20 and 0.30 were added back into the test, and to then decline to 0.735 as the weaker items were added back into the test. The point is that essentially only one third of the items (those with item-to-total score correlation's in excess of 0.30) were providing the discrimination among the test takers. The other two-thirds of the items, taken as a group, provided no additional discrimination nor additional reliability. They did serve to produce a longer test, but they introduced additional error into the test scores, and their contribution to the test's validity is, therefore, questionable at best.

In our judgment, none of the 27 tests that we analyzed could or should be considered to be valid tests of the content domains or constructs that they purportedly measured. They all contain far too many "weak" items and far too few discriminating items. They are invalid not because the items did not pertain appropriately to the content domain, but because many of the items did not contribute enough to the valid measurement of the content domain. Indeed, about 1/2 of the items in each of the tests contributed variance to the total scores that might best be characterized as essentially error variance.

It is possible that we are being too harsh. Perhaps we should say that the test developer who created the test depicted in Figure I is off to a good start. After all, substantial improvement is possible if the worst 25 items are rewritten or replaced. Inspection of Figure 1 suggests that improvement is possible to the extent that coefficient alpha could start on the left as high as 0.90 and the parabola could open downward monotonically from there. It would appear that with revisions, the test could produce an alpha as high as 0.900 or so. Having already achieved an alpha of 0.814 with the best 1/2 of the items, surely the developer can revise the remaining items and at the very least retain an alpha of 0.814 for the whole test, if not cause it be increased. Indeed, applying the Spearman-Brown formula to the half test consisting of the best 25 items with a reliability of 0.814 indicates that an additional 25 comparable items would yield a reliability coefficient for the whole test of approximately 0.90.

As a final note, we would suggest that in this paper we have analyzed the effect that a weak item, or a collection of weak items, has on the reliability of a classroom test as depicted by coefficient alpha. In each of the 27 tests that we analyzed, the effect was substantial, and much improvement in the reliabilities of these tests was possible. Our focus has been on improving the test. We are not advocating deleting items to improve reliability, nor are we particularly interested in improving reliability for its own sake. On the assumption that the test constructor writes or revises items that are appropriate replacements

related to the construct or content domain, then improving individual items' abilities to discriminate among test takers will improve the validity of the test. On the other hand, even with "very appropriate" items, a test like the one depicted in Figure 1 with a coefficient alpha of only 0.735 indicating it measures knowledge of the content domain with a great deal of error, is not a valid test of knowledge.

REFERENCES

Hopkins, K. D. (1998). *Educational and Psychological Measurement and Evaluation* (8th ed.). Boston: Allyn and Bacon.

Sax, Gilbert. (1997). *Principles of Educational Psychological Measurement and Evaluation* (4th ed.). Belmont, CA: Wadsworth Publishing Co.

Thorndike, Robert M. (1997). *Measurement and Evaluation in Psychology and Education* (6th ed.). Upper Saddle River, NJ: Prentice-Hall, Inc.

Dale Shaw is a professor in the College of Education at the University of Northern Colorado. He teaches research methods and statistics courses in the Department of Applied Statistics and Research Methods. His research focuses on measurement theory, online teaching, and teaching effectiveness.

Suzanne Young is an associate professor in the College of Education at the University of Wyoming. She teaches research methods, statistics, and measurement courses. Her research interests include teaching effectiveness, online teaching and learning, and measurement theory.

We would like to acknowledge the contribution of Alan Harney who wrote the computer program for the specialized iterative item analysis