DOCUMENT RESUME

ED 319 795 TM 015 074

AUTHOR Trevisan, Michael S.; Sax, Gilbert
TITLE Reliability and Validity of Multiple-Choice Examinations as a Function of the Number of Options per Item and Student Ability.
PUB DATE Apr 90
NOTE 37p.; Paper presented at the Annual Meeting of the American Educational Research Association (Boston, MA, April 16-20, 1990).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Ability Grouping; *Academic Ability; *College Bound Students; High Achievement; High Schools; *High School Students; Low Achievement; *Multiple Choice Tests; Parochial Schools; Test Construction; Test Format; Test Items; Test Reliability; *Test Validity; Track System (Education); Verbal Tests
IDENTIFIERS *Washington Precollege Testing Program

ABSTRACT

Reliability and validity of multiple-choice examinations as a function of the number of options per item and of student ability were computed for 435 junior class parochial high school students in the tri-county area of Portland (Oregon). The verbal section of the Washington Pre-College Test Battery was used. The least discriminating options were deleted to create 3-option and 4-option test formats from the original 5-option per item test. Students were placed into ability groups using grade point average (GPA) cutoffs. The GPAs were the criterion for the validity coefficients. Significant differences were found between reliability coefficients for 3- and 5-option formats for low ability students. Non-significant differences were found between reliability coefficients for high or average ability students. The optimum number of options across ability groups was three. None of the validity coefficients followed the trend expected. The results are part of the mounting evidence suggesting the efficacy of the -option item. Five tables present study data. (Author/SLD)

RELIABILITY AND VALIDITY OF MULTIPLE-CHOICE EXAMINATIONS AS A FUNCTION OF THE NUMBER OF OPTIONS PER ITEM AND STUDENT ABILITY

by

Michael S. Trevisan and Gilbert Sax
University of Washington

Paper presented at the American Educational Research Association Conference, Spring, 1990, Boston, Massachusetts, Division D

ABSTRACT

Reliability and validity of multiple-choice examinations as a function of the number of options per item and student ability was computed for junior class parochial high school students in the tri-county area of Portland, Oregon. The verbal section of the Washington Pre-College Test Battery was used. The least discriminating options were deleted to create 3- and 4-option test formats from the original 5-option per item test. Students were placed into ability groups using grade point average (GPA) cutoffs. The GPAs were the criterion for the validity coefficients. Significant differences ($p \leq 0.05$) were found between reliability coefficients for low ability students; however, the trend favored the 4-, 3-, and 5-option formats, respectively, which was not the trend hypothesized.

Nonsignificant differences were found between reliability coefficients for high or average ability students. The optimum number of options across ability groups was three. None of th validity coefficients followed the trend expected. These results are part of the mounting evidence which suggest the efficacy of the 3-option item. An explanation is provided.

Much of the theoretical literature which has investigated the reliability of multiple-choice examinations as a function of the number of options per item suggest the efficacy of the 3-option item (e.g., Ebel, 1969; Grier, 1975; Lord, 1944). These studies however, assume knowledge or random guessing when a subject is confronted with an item. Recognizing that student ability may play a role in the efficacy of the item format, Lord (1977) scaled item response data from the College Board Scholastic Aptitude Test with the 3-parameter logistic function. Lord's results showed a more efficient test for high ability students as the number of options per item decreased but less efficient for low ability examinees. Also, for the middle 80% of the ability range, the 3-option item provided the most efficient test. An assumption Lord made in this study was that testing time is proportional to test length. This is known as the Tversky condition (Tversky, 1964).

Several empirical studies have also suggested the efficacy of the 3-option item over the 4- or 5-option item (e.g., Costin, 1970,1972; Owen and Froman, 1987; Straton and Catts, 1980). These studies however, did not consider the interaction between the number of options

per item and student ability and the effects this might have on the reliability of the test.

Only three empirical studies have considered the interaction between student ability and the number of options per item as suggested by Lord (1977).

Weber (1978) provided two reports, first comparing the reliabilities of 3- vs. 5-option tests, and second comparing the reliabilities of 3- vs. 4-option tests. Weber (1978) concluded that for low achieving students reliability estimates were improved as the number of options per item was increased and that for high achieving students no such gains were found.

Although this study did not corroborate the findings from Lord (1977), several limitations in this study may have influenced the results. First, the different versions of the test were administered with only a short time interval between them. The result was a repeated measures situation with memory and practice effects influencing the results of the second testing. Second, the means of classifying students into ability groups was by district personnel whose opinions might not be wholly reliable and valid. Third, these tests were very short and the N's very small which would greatly influence the

magnitude of the reliability coefficients. Last, the Tversky condition was not invoked in either study.

In an attempt to improve upon the design limitations of Weber (1978), Green, Sax, and Michael (1982) compared the reliabilities of 3-, 4-, and 5-option tests. Nine sections of a beginning French class were used in this study. Teacher judgment was used to eliminate options. The Tversky condition was invoked by using the Spearman-Brown formula to predict what the reliability would have been if the length of the 3- and 4-option test formats was increased to encompass the total number of options in the 5-option format test. Tests were administered randomly by section. Students were grouped into ability groups using final course grade points in the following manner: high (3.6-4.0), average (3.1-3.5), and low (0-3.0). Final course grades were also used as the criterion in calculating the validity coefficients for each test. According to Green, Sax, and Michael (1982), the results of the study were:

> Differences among reliabilities for the low ability group for the 3-, 4-, and 5-choice tests were significant at alpha=.05 (M=10, df=2) but the trend of the reliabilities was clearly not the one indicated by Hypothesis 1. In decreasing order of magnitude, KR- 20's favored the 4-choice test, the 3-choice test, and 5-choice test. For the high ability group, differences among reliabilities were not

> significant; for the average ability group, differences were significant at alpha=.10 (M=5.94, df=2). The latter two results were contrary to Hypotheses 2 and 3. After the ability groups were combined differences among reliabilities for the 3-, 4-, and 5-option tests were compared. These differences were significant at alpha=.05 (M=13.73, df=2). The optimal number of alternatives for all ability groups combined was four. (p.242)

The results also showed that none of the validity coefficients for prediction of final course grade was significantly different. However, predictions tended to be most accurate for the high ability group.

Differences between this study and those of Lord (1977) and Weber (1978) may be attributed to several factors. First, the overall item p-value was 0.78 while those in the Lord (1977) and Weber (1978) studies were 0.50 and 0.67, respectively. Most students in the Green, Sax, & Michael study were classified in the higher ability range based on the GPA cutoffs. Consequently, this was a fairly easy test given to a high ability group. This combination will attenuate the reliability coefficients. Tests were randomly adminstered by class. If tests were randomly administered by student rather than by class, greater control of systematic error would be achieved because the error term in the test statistic would be based on a greater number of degrees of freedom.

This would reduce the magnitude of the error term and increase the magnitude of the test statistic. Also, the contiguous nature of the cutoff scores to delineate the ability groups could blur the differences between them and, therefore, could decrease statistical power. Last, the distractor elimination technique was by teacher judgment. Perhaps higher reliability estimates might have been obtained if options could have been eliminated using discrimination indices, a practice recommended by several authors (e.g., Costin, 1970; Lord, 1977; Williams and Ebel, 1957).

Despite these differences in results, this study provided information that could be extended for further study of the interaction of student ability and test format for classroom tests. More specifically, the assignment of tests on a random basis and the use of a test statistic provided experimental control that did not exist in previous studies. Also, more reliable ability estimates were employed than was used in the Weber (1978) study. And last, evidence for validity was provided which has been virtually nonexistent in previous studies.

A third study which considered the student ability by item format was conducted by Levine and Drasgow (1982). A total of 9900 examinees from the

April, 1973 Graduate Record Examination-Verbal section (GRE-V) and 75,000 examinees from the April, 1975 Scholastic Aptitude Test-Verbal section (SAT-V) provided the responses for this study.

Response patterns of the examinees for each option in the items was analyzed to determine if an interaction between student ability and the options chosen could be detected. Students were divided into ability groups using deciles within the score distribution for each test. Histograms were constructed to examine the pattern of option selection for examinees at various ability levels. By examining the response patterns for those examinees who answered the item incorrectly, implications were made concerning the appropriate number of options per item necessary to maximize information from an item for a particular ability group. For high ability students, one or two distractors were generally chosen for many of the items. As the ability level decreased, there was a tendency for more of the distractors to be chosen by these ability groups. These results suggest that information will be maximized by using more options per item for lower ability groups and lower numbers of options for higher ability groups.

The Tversky condition was not implemented in this study. Constructing histograms also provided a means of studying the student ability by item format interaction without employing an option-elimination technique. And although this was not a classroom study as were the previous two, empirical evidence was provided which corroborates the findings of Lord (1977).

Given the previous findings then, it is the purpose of the present study to compare the reliabilities and validities of 3-, 4-, and 5-choice tests for examinees of different abilities. It is an attempt to provide empirical evidence for Lord's findings concerning the student ability by test format interaction. It provides an improvement and extension of the Green, Sax, & Michael (1982) design by (a) randomizing the tests by student rather than by se tion which will further reduce systematic error and increase the power of the design to detect treatment effects; (b) increase the power of the test statistic by spreading out the distance between ability groups on the range of GPA scores and decreasing the standard deviation within groups; (c) eliminate the least discriminating options when creating the 3- and 4-option formats which will systematically increase reliability in these tests; a : (d) use items with p-

values that more closely approximate those used by Lord (1977).

Based on the findings of Green, Sax, and Michael (1982), the hypotheses of this study are:

1. Significant differences exist between internal consistency reliability coefficients (Kuder-Richardson 20) for low-ability students and will favor the 5-, 4-, and 3-option tests, respectively.

2. No significant differences exist between reliability coefficients as number of options decreases for high-ability level students ($p \leq 0.15$). (Note: This is a research hypothesis stated in null form. The required significance level will be increased to 0.15 to reduce the probability of a Type II error.)

3. Significant differences exist between reliability estimates for average-ability students and will favor the 4-, 5-, and 3- choice tests, respectively. (These were the findings of Green, Sax, & Michael (1982) which were significant at alpha = 0.10. Therefore, this hypothesis is an attempt to corroborate those findings.)

4. The optimum number of alternatives for all ability groups combined is four.

5. The validity coefficients will follow the same pattern as hypothesized for the reliability coefficients for each ability group.

METHOD

A total of 435 junior class parochial high school students from the tri-county area of Portland, Oregon participated in this study. The verbal subsection from the Washington Pre-College Admissions Test Battery was given, which is a 5-option multiple-choice examination normed with a random sample of junior class high school students from the state of Washington. To construct the 4- and 3-choice test formats distractors were eliminated by using the point biserials provided with the standardization data and eliminating the least discriminating options.

Students were given 12 minutes to complete the 45 item test which was the time requirement given in the administration manual. Since 98% of the students finished the test in the allotted time, speed was not a significant factor in their performance.

The three forms of the examination were randomly assigned to each student. Students were also asked to estimate their current GPA and mark this on their answer sheet.

Students were later grouped into high, medium, and low ability groups using the following GPA cutoffs: high (3.6 -4.00), medium (3.00 - 3.2), and low (0.00 - 2.6). Students with GPAs between these ability group cutoffs were not used in this part of the analysis. It was not possible to verify the students' estimate of their GPA . However, a check on the validity of these estimates was accomplished by conducting an analysis of variance on the mean GPA estimates for the three ability groups within each test format. Significant differences were found ($p \leq 0.05$) in all cases. Further evidence for the accuracy of the GPA estimates was established by applying the Newman-Keuls multiple-comparison technique. All pairwise comparisons were also significant ($p \leq 0.05$); thus, establishing the existence of a trend among the means which would be reflective of the trend in the ability groups. One final check on the validity of the GPA estimates was accomplished by conducting an analysis of variance on the mean GPA estimates when the ability groups are combined for each form of the test.

Nonsignificant differences were found ($p \leq 0.05$) among the combined groups mean GPA estimates for the forms of the test, providing evidence that the range of ability estimates within each test format were comparable, as would be expected. The GPA estimates served as the criterion for the validity coefficients.

RESULTS

Table 1 presents the means, standard deviations, sample sizes for each form of the test and each ability group, as well as the mean p-values, the number of items with nonzero variances and the mean GPAs.

KR-20s were calculated for each test form and ability group. The Tversky condition (Tversky, 1964), which assumes that testing time is proportional to test length was implemented by using the Spearman-Brown formula to determine the reliability for a lengthened version of the 3- and 4-option test forms given the total number of options found in the 5-option form.

KR-20s were computed for each test form and ability group. The KR-20s were statistically compared using the M statistic (Hakstian and Whalen, 1976). The statistic is distributed as chi-square, conforms to the requirements of a two-factor random effects model of the

analysis of variance, and provides a test of the null hypothesis that reliability coefficients from independent samples are equal.

Correlations between test scores and GPAs were also computed (validity coefficients) and statistically compared using the U statistic (Marascuilo, 1966). This statistic is also distributed as chi-square and provides a test of the null hypothesis that validity coefficients from independent samples are equal.

Table 2 presents KR-20s for the 45 item forms of the test and the KR-20s for the forms of the test which were lengthened to invoke the Tversky condition. These interal consistency estimates are referred to as the unadjusted and adjusted KR-20s, respectively. Also presented in Table 2 are the standard errors of measurement and validity coefficients for each form of the test and ability group.

(INSERT TABLE 1 ABOUT HERE)

(INSERT TABLE 2 ABOUT HERE)

The results of the study showed that significant differences exist ($p \leq 0.05$) between reliability coefficients for low ability students with a chi-square equal to 9.21 (df=2); however, the trend favored the 4-,

3-, and 5-option format, which was not the trend hypothesized. Nonsignificant differences were found between reliability coefficients for high ability (χ^2=0.29, df=2) or average ability (χ^2=2.97, df=2) students. The optimum number of options across ability groups was three. None of the validity coefficients followed the trend expected.

DISCUSSION

Partial explanation of these findings can be gained by considering previous empirical studies which considered the option by student ability interaction. Weber (1978) found little difference among reliability coefficients (although not compared statistically) of the different item formats for high ability students, corroborating the current findings, but found the reliability to increase for low achieving students as the number of options increased, which does not corroborate the current findings. Green, Sax, and Michael (1982) found no significant differences among reliability coefficients for the high ability group as did the current study. However, the current study corroborated one finding, validated by Green, Sax, and Michael (1982); namely, that the 5-option format is the least defensible.

When considering the item elimination process in this study and the effects this process had on the results, the point biserials of the items and distractors were examined and categorized in several ways which help describe their influence.

One way that this was done was to examine the point biserials from the standarization data. Twenty-five of the items from the original 5-option per item test had distractors with point biserials within 0.05 in magnitude from one another. Consequently, little was gained in discrimination between the 3- and 4-option test forms. This may have contributed to some of the nonsignificant findings.

Another way in which the items and distractors were examined was to compare the magnitude of the point biserials for a single item or distractor across test forms. No pattern could be detected among the point biserials for a single item across test forms within any of the ability groups. Also, no pattern could be detected among the point biserials for a single distractor across test forms within any of the ability groups. Consequently, the influence the option elimination process had on a single item or distractor is not known.

However, by classifying the item point biserials with the Educational Assessment Center classification system at the University of Washington (University of Washington, 1989), several trends among the item point biserials were detected. (The categories are defined as follows: A poor discrimination index was defined as a point-biserial that falls in the range between less than zero to 0.09; fair, any value between 0.10 to 0.29; and good, values equal to or greater than 0.30.)

First, for all test forms and ability groups, the number of items which were categorized as good discriminators ranged from a low of 18 for the 5-option form for the low ability group to a high of 30 for the 3-option form when the ability groups were combined. The number of items which were classified as fair discriminators ranged from a low of 9 for the 5-option test form when the ability groups were combined to a high of 20 for the 5-option test form given to the low ability group. Poor discriminators ranged from 0 for all test forms when the ability groups were combined to 10 for the 4-option form given to the low ability group. A general trend which can be gleaned from this categorization is that the number of items classified as good discriminators increased as the ability level increased

and the number of options per item decreased. Examined in this way, the item point biserials indicate that the 3-option form has high discriminating items for the high ability group. This however, was also true for the average and low ability groups, but the number of items classified as good dscriminators was not as great as it was for the high ability group. For the combined ability groups, more items were classified as good discriminators for all test forms.

A different way of categorizing the items was to determine the number of items with one or more distractors not chosen, indicating dysfunctional distractors. This was done for each ability group and test form and can be found in Table 3. The number of items with dysfunctional distractors ranged from a low of 1 for the 3-option form for the low ability group to a high of 28 for the 5-option form given to the high ability group. For all ability groups, the number of items with dysfunctional distractors decreased as the number of options decreased. However, when the ability groups were combined, no test form had any items with dysfunctional distractors except for the 5-option form which had 4 items.

(INSERT TABLE 3 ABOUT HERE)

The last way in which the items were classified was to determine the number of items which had point biserials smaller in magnitude than one or more of its distractors. This was done for each form and ability group and can be found in Table 4. The number of items with point biserial smaller in magnitude than one or more distractors ranged from a low of 0 in the 3-option form for the low and average ability groups to a high of 10 for the 5-option form given to the low ability group. In general, the number of items with point biserials smaller in magnitude than one or more of its distractors decreased as the number of options per item decreased for all ability groups. For the combined ability groups, all forms had 0 items with point biserials smaller in magnitude than one or more of its distractors except the 5-option form which had only 1.

(INSERT TABLE 4 ABOUT HERE)

While these categorizations may not directly explain the reliability results, all suggest the efficacy of the 3-option test form over any of the other forms. This is true for all ability groups and also when the ability groups are combined. The 3-option test form has more items that are high discriminators than do the 4- or 5-

option test forms. The 3-option form also has less items with dysfuncitonal distractors or point biserials smaller in magnitude than one or more of its distractors than did the 4- or 5-option test forms.

Further explanation of the findings for the reliability coefficients may be obtained by considering the sample of students used for the standardization data of the test and the time it was standardized. The standardization data were obtained from an October, 1983 administration of this test. The mean score was 13.18 with a standard deviation of 11.53 from a random sample of eleventh grade students across the state of Washington, a small proportion of whom were parochial school students (R. Greenmun, personal communication, January 25, 1990). The KR-20 was 0.95. In the current study, eleventh grade parochial school students from Northwest Oregon were tested in September, 1989. The combined groups mean test scores for the test formats ranged from 21.64 to 25.87, considerably higher than the standardization group. The range of standard deviations for the current study ranged from a low of 7.92 to a high of 8.65, which is considerably lower than the standardization group. The KR-20s in this study ranged from a low of 0.73 to a high of 0.92. Although these

KR-20s were adjusted for the Tversky condition and consequently, higher in magnitude than if they were not adjusted, they are lower in magnitude than the KR-20 obtained from the standardization group of examinees. These results point to the possibility that the magnitude of the reliability coefficients in this study may have been attenuated by the restricted population of students who were tested. A further study with a student group more reflective of the norm group is warranted.

The time difference (6 years) from when the test used in this study was normed to the time the test was given for purposes of this study, may also account for the lower reliability coefficients. Normative data on achievement tests is time-bound and reflective of the students who took the test for norming purposes (Williams, 1988). National achievement indicators point to an increase in achievement since the late 1970's and early 1980's (Williams, 1988). A general increase in achievement over the last six years would restrict the range of achievement of students used in this study as compared to the norm group, since students who participated in this study are generally higher achievers than in the norm group as evidenced by the group means. This restriction in range would attenuate the magnitude

of the reliability coefficients. A further study would warrant the use of current standardization data.

The method of ability grouping may have also affected the results. The GPA cutoffs were established to give the greatest number of subjects in each ability group and test form given the original number of subjects tested, while attempting to minimize the within group variance relative to any between group variance which may have been present. It is difficult from the present data to determine whether or not greater or lesser control of the within group variability would affect the results. Further study regarding the appropriateness of a contiguous or noncontiguous design is warranted.

The validity findings are difficult to explain. These did not corroborate the findings of the Green, Sax, and Michael (1982) study which showed that prediction of final course grade was most accurate for the high ability group. Perhaps partial explanation for the low magnitude of the validity coefficients can be attributed to the restricted range of students who participated in this study. The restricted range of students in this study may have attenuated the validity coefficients and may account for some of the nonsignificant findings.

It is also difficult to assess the extent to which the method of ability estimation influenced the results. While validity evidence was established by considering between group and format differences as previously mentioned, it cannot be determined from these data whether or not students, as a whole, under- or over-estimated their GPAs. A systematic bias of GPA estimation in one direction may affect the trend of the validity coefficients as well as attenuate their magnitude. A further study using a standardized ability measure may alleviate this possible problem.

As with the reliability findings, the method of ability grouping may have also affected the results. Further study regarding the noncontiguous nature of this design for purposes of validity is also warranted.

ADDITIONAL ANALYSES

To address some of the concerns previously mentioned, further analyses of the data were conducted. This was done in three ways. First, the significance of the reliability coefficients was tested after the average ability group was eliminated for each of the test formats. This would increase the distance between the low and high ability groups and provide some evidence as

to whether or not the GPA cutoffs used had any affect on the reliability results. Second, a contiguous design was developed and used. This was done by establishing a cutoff at a GPA of 3.00 between the low and high ability groups. Subjects eliminated in the original study to establish the noncontiguous ability groups were added to this part of the study to establish the contiguous ability groups. The results of this part of the study were contrasted with the reliability findings when the average ability group was eliminated. If nonsignifcant differences are found among reliability coefficients in the contiguous design and significant findings were found in the noncontiguous design, this would be evidence for the usefulness of the noncontiguous design when comparing the reliabilities among ability groups and item formats. Third, the 4-option format was eliminated to contrast the 3- and 5-option formats for each of the ability groups. This increases the difference between the number of options per item compared. If nonsignificant differences are found, more evidence for the usefulness of the 3-option item would be obtained. The Tversky condition was invoked for the additional reliability comparisons. Validity coefficients were calculated and compared.

RESULTS

Means, standard deviations, sample sizes for the first and third reliability comparisons were previousl calculated and are located in Table 1. Also, the reliability coefficients and standard errors of measurement as well as the adjusted versions of each of these (adjusted to invoke the Tversky condition) have also been previously computed and are located in Table 2 along with the validity coefficients. For the second reliability comparison, the aforementioned descriptive statistics can be found in Table 5.

(INSERT TABLE 5 ABOUT HERE)

Reliability findings

For the noncontiguous comparisons significant differences were found ($p \leq 0.05$) between the high and low ability groups for the 5- and 4-option formats with chi-squares of 7.42 (df=1) and 5.89 (df=1), respectively; however, nonsignificant differences were found between the high and low ability groups for the 3-option format.

The results of the contiguous comparison showed that nonsignificant differences exist between the high and low ability groups for all test formats.

When the reliability estimates between the 3- and 5-option formats were compared, significant differences were found ($p \leq 0.05$) for the low ability group with a chi-square of 2.77 (df=1). Nonsignificant differences were found between the 3- and 5-option formats for the average and high ability groups.

Contrasting the findings of contiguous and noncontiguous comparisons suggests the usefulness of a noncontiguous design when comparing reliability coefficients from different ability groups and item formats. What remains for further study is to determine the optimum cutoff values along the ability continuum to maximize the power of the design to detect differences among the reliability coefficients.

The findings for the last comparison again suggest the efficacy and efficiency of the 3-option format over the 5-option format, regardless of the ability group.

Validity Findings

Nonsignificant differences were found for all the validity comparisons which correspond with the additional reliability comparisons ($p \leq 0.05$).

Changes made in the additional analyses had little effect on the validity results, since nonsignificant

differences between validity coefficients were found in all cases. In general however, the magnitude of the validity coefficients increased when the sample size was increased in the contiguous comparison, ranging in value from a low of 0.09 for the low ability group given the 3-option test to a high of 0.43 for the high ability group given the 5-option test. Further validity study is needed using larger sample sizes than were used in the original part of this study.

Theoretically, (for example, see Lord, 1977), this study did not corroborate a negative reliability coefficient between the number of options per item and ability level, when it is assumed that testing time is proportional to test length. Also, this study did not corroborate previous empirical findings with regard to differences in reliability estimates when considering the interaction between student ability and the number of options per item (i.e., Weber, 1978; Green, Sax, and Michael, 1982). And although the results of this study do not confirm research expections, except the research expectations of the reliability estimates for the high ability group, this study is part of the mounting evidence which show little difference between the reliability estimates of different test formats. In

practice, these results add evidence to the argument provided by Costin (1970;1972) and Ebel (1969) concerning the advantage of constructing 3-option items over 4- or 5-option items. Specifically, since similar reliability estimates may be obtained when using any of the aforementioned item formats, considerable time and energy will be saved if 3-option items are constructed, rather than constructing 4- or 5-option items.

REFERENCES

Allen, M. J. & Yen, W. M. (1979). Introduction to Measurement Theory. Monterey, CA: Brooks/Cole Publishing Co.

Budesco, D. & Nevo, B. (1985). Optimal number of options: an investigation of the assumption of proportionality. Journal of Educational Measurement, 22, 183-196.

Costin, F. (1970). The optimal number of alternatives in multiple-choice tests: Some empirical evidence for a mathematical proof. Educational and Psychological Measurement, 30, 353-358.

Costin, F. (1972). Three-choice versus four-choice items: Implications for reliability and validity of objective achievement tests. Educational and Psychological Measurement, 32, 1035-1038.

Ebel, R. (1969). Expected reliability as a function of choices per item. Educational and Psychological Measurement, 29, 565-570.

Educational Testing Service. College Board Scholastic Aptitude Test. Princeton, N.J., 1947-1987

Grier, J. B. (1975). The number of alternatives for optimum test reliability, Journal of Educational Measurement, 12, 109-113.

Green, K., Sax, G., & Michael, W. B. (1982). Validity and reliability of tests having differing numbers of options for students of differing levels of ability. Educational and Psychological Measurement, 42, 239-245.

Hakstian, A. R. & Whalen, T. E. (1976). A k-sample significance test for independent alpha coefficients. Psychometrika, 41, 219-231.

Haladyna, T. M. (1989). Validity of a taxonomy of multiple-choice item-writing rules. Applied Measurement in Education, 2, 51-78.

Levine, M. V. & Drasgow, F. (1983). The relation between incorrect option choice and estimated ability. Educational and Psychological Measurement, 43, 675-685.

Lord, F. M. (1977). Optimal number of choices per item-A comparison of four approaches. Journal of Educational Measurement, 14, 33-38.

Marascuilo, L. (1966). Large-sample multiple comparisons. Psychological Bulletin, 65, 280-290.

Owen, S. V. & Froman, R. D. (1987). What's wrong with three-option multiple choice items? Educational and Psychological Measurement, 47, 513-522.

Stiggins, R. (1988). Revitalizing classroom assessment: the highest instructional priority. Phi Delta Kappan, 69(5), pp. 363-368.

Straton, R. G. & Catts, R. M. (1980). A comparison of two, three, four-choice item tests given a fixed total number of choices. Educational and Psychological Measurement, 40, 357-365.

Toops, H. A. (1921). Trade tests in education. Teachers College Contributions to Education, #115. N.Y.: Columbia University, Teachers College.

Tversky, A. (1964). On the optimal number of alternatives at a choice point. Journal of Mathematical Psychology, 1, 386-391.

University of Washington (1989). Scorepak: Item Analysis. Educational Assessment Center.

Weber, M. B. (1978). The Effect of Choice Format on Internal Consistency, Paper presented at the National Council on Measurement in Education Annual Meeting, Toronto, Canada.

Williams, B. J. & Ebel, R. L. (1957). The effect of varying the number of alternatives per item on multiple-choice vocabulary test items. The Fourteenth Yearbook, National Council on Measurements Used in Education, pp. 63-65.

Williams, P. L. (1988). The Time-bound nature of norms: understandings and misunderstandings. Educational Measurement: Issues and Practices, 7, 18-21.

TABLE 1

MEANS, STANDARD DEVIATIONS, SAMPLE SIZES, MEAN GPAS, MEAN p-VALUES AND NUMBER OF ITEMS WITH NONZERO VARIANCES FOR EACH TEST FORM AND ABILITY GROUP

TEST FORM	p	ITEMS (s^2=0)	X	s	SAMPLE SIZE	MEAN GPA
3-OPTION						
L	0.51	45	23.19	7.71	32	2.31
A	0.57	44	25.90	8.18	31	3.05
H	0.63	45	28.35	7.37	34	3.82
C	0.57	45	25.87	8.04	97	3.07
4-OPTION						
L	0.38	45	17.16	5.21	25	2.29
A	0.46	45	21.09	7.57	34	3.07
H	0.58	43	26.07	7.86	30	3.79
C	0.48	45	21.66	7.92	89	3.09
5-OPTION						
L	0.38	45	17.46	5.76	35	2.22
A	0.46	45	20.67	7.65	30	3.08
H	0.61	45	27.29	9.23	31	3.79
C	0.48	45	21.64	8.65	96	3.00

Note: L = Low Ability A = Average Ability
H= High Ability C = Combined Ability Groups

TABLE 2

UNADJUSTED AND ADJUSTED KR-20s, STANDARD ERRORS OF MEASUREMENT AND VALIDITY COEFFICIENTS FOR EACH TEST FORM AND ABILITY GROUP

TEST FORM	UNADJ. KR-20	ADJ. KR-20	UNADJ. SEM	ADJ. SEM	VALIDITY COEFFICIENT
3-OPTION					
L	0.85	0.90	2.98	2.43	-0.13*
A	0.87	0.92	2.90	2.31	0.04
H	0.85	0.90	2.89	2.33	0.01
C	0.87	0.92	2.95	2.27	0.24
4-OPTION					
L	0.68	0.73	2.93	2.70	0.05*
A	0.85	0.88	2.93	2.62	0.36
H	0.87	0.89	2.83	2.60	0.37
C	0.86	0.88	2.93	2.74	0.47
5-OPTION					
L	0.75	0.75	2.90	2.90	-0.14*
A	0.85	0.85	2.93	2.93	-0.13
H	0.91	0.91	2.73	2.73	-0.19
C	0.89	0.89	2.89	2.89	0.42

Note: L = Low Ability A = Average Ability
H = High Ability C = Combined Ability Groups

TABLE 3

Items with 1, 2, or 3 distractors not chosen

Ability

Form	Low	Average	High	Combined
3-option	1,4,9	1,6,9,14	1,6,9,11	NONE
4-option	1,9	4,6,9,14	1,3,4,6, 7,8,9,11, 14,18,27, 31,37,38, 44	NONE
5-option	1,2,4, 5,7,9, 17,26,	1,3,4,8, 9,10,13, 15,18,26, 28,31,37, 38,42,45	1,2,3,4, 6,7,8,9, 10,11,12, 13,14,17, 18,19,22, 23,25,28, 30,31,32, 37,38,43, 45	NONE

TABLE 4

Items with point biserials smaller in magnitude than than one or more of its distractor

Form	Low	Average	High	Combined
3-option	NONE	NONE	44	NONE
4-option	10,43	36,41	26,30	NONE
5-option	1,5,14 15,26,38, 40,41,43,	1,8,10 17,34,45	14	NONE

TABLE 5

MEANS, STANDARD DEVIATIONS, SAMPLE SIZES, ADJUSTED AND UNADJUSTED RELIABILITY COEFFICIENTS AND STANDARD ERRORS OF MEASUREMENT, AND VALIDITY COEFFICIENTS FOR THE NONCONTIGUOUS COMPARISON

TEST FORM	X	s	SMP. SIZE	UN. KR-20	ADJ. KR-20	UN. SEM	ADJ. SEM	r
3-OPT.								
L	24.36	7.94	75	0.86	0.91	2.97	2.97	0.09
H	26.22	7.54	73	0.85	0.90	2.94	2.38	0.24
4-OPT.								
L	18.13	6.49	64	0.79	0.84	2.98	2.59	0.10
H	24.35	7.78	71	0.86	0.89	2.90	2.58	0.28
5-OPT.								
L	19.03	7.68	71	0.86	0.86	2.90	2.90	0.15
H	21.82	9.37	77	0.91	0.91	2.82	2.82	0.43

Note: OPT.=option; SMP.=sample; UN.=unadjusted;

ADJ.=adjusted