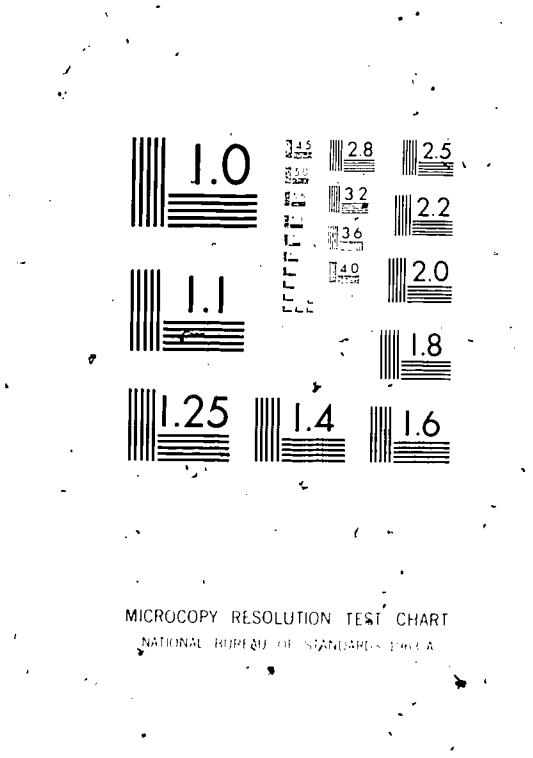

1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963-A

DOCUMENT RESUME

ED 169 073 TM 007 865

AUTHOR Green, Kathy
TITLE Multiple Choice Converted to True-False: Comparative Reliabilities and Validities.
PUB DATE 78
NOTE 10p.; Paper presented at the Annual Meeting of the Western Psychological Association (San Francisco, California, 1978)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Tests; *Comparative Testing; Higher Education; *Multiple Choice Tests; Reaction Time; Response Style (Tests); *Test Construction; *Test Reliability; *Test Validity
IDENTIFIERS Teacher Developed Tests; *True False Tests

ABSTRACT

Forty three-option multiple choice (MC) statements on a midterm examination were converted to 120 true-false (TF) statements, identical in content. Test forms (MC and TF) were randomly administered to 50 undergraduates, to investigate the validity and internal consistency reliability of the two forms. A Kuder-Richardson formula 20 reliability was computed for each form. Reliability of the MC form was then adjusted with the Spearman-Brown formula to equate testing time, since the MC form took three-fourths as much time to complete as the TF form. Adjusted reliability coefficients of the TF and MC forms were .80 and .73, respectively. To compare validity, a Pearson product moment correlation was computed between test score and grade point average; validity coefficients were .49 (TF) and .52 (MC). Results support the use of TF teacher made tests as alternatives to MC tests with no loss in reliability or validity. However, as previous studies have shown, these results are not obtained when MC items are revised and the range of the TF form is restricted. (CP)

MULTIPLE CHOICE CONVERTED TO TRUE-FALSE: COMPARATIVE RELIABILITIES AND VALIDITIES

Kathy Green

Paper presented at the Western Psychological Association Annual Convention, 1978, San Francisco.

The advantages and disadvantages of multiple-choice and true-false formats have been studied by a number of investigators (e.g. Nunnally, 1964; Karmel, 1970; Blood & Budd, 1972). One advantage claimed for the true-false (TF) test is that it allows a more efficient sampling of course content. Proponents of the multiple-choice (MC) format, however, argue that this advantage may be offset by lowered reliability coefficients that occur primarily as the result of guessing on TF items. Evidence is available to support the contention that reliability (Oosterhof & Glasnapp, 1974) if not validity (Frisbie, 1973, 1974) of the MC test form is higher than that of the TF test form.

Empirical comparisons of the TF and MC formats, however, have been beset by a number of methodological problems. One serious difficulty is that greater care is often given to the preparation of MC items than to TF items. For example, more MC than TF tests have been item analyzed and revised prior to their being compared (Frisbie, 1973, 1974; Oosterhof & Glasnapp, 1974). It would be expected that if MC items were more extensively revised than TF items, the reliability of the MC form would be higher. The present study employed teacher-made tests in which the MC and TF items were not differentially improved.

To further complicate matters, TF items are often constructed from parallel MC items on either a one-to-one or on a two-to-one basis--either one or two questions being generated from each MC item. Resultant test forms then have contained as many or twice as many TF as MC items. However, even at a ratio of 2:1 TF:MC questions, an estimate of the hypothetically lengthened TF form reliability was necessary to equate testing time (Oosterhof & Glasnapp, 1974).

Frisbie (1973) suggested that use of a longer TF test would probably increase the variance of the TF scores and produce a better estimate of the relationship between MC and TF forms. The present study converted each three-option MC statement to three separate parallel TF statements, obviating the need to adjust the reliability of the TF form to equate testing time and providing for increased variance of the TF scores.

The expectations for this study were:

1. When MC items are converted to TF items, the internal consistency reliabilities of the two forms do not differ significantly.
2. When testing time is equated, reliabilities of the two forms do not differ significantly.
3. Validity of the two forms does not differ significantly.

To reduce likelihood of Type II errors, null hypotheses required rejection at the .10 rather than the .05 level.

METHOD

Subjects

This study was conducted during the summer quarter of 1977. Fifty undergraduates enrolled in a required introductory class in tests and measurement at the University of Washington served as subjects. Subjects were naive at the time of testing regarding the nature of reliability and the relationship between MC and TF item formats.

Instrumentation

A MC and TF form of a midterm examination were constructed for the class. Items on the two forms differed in format only, the content being identical in every instance. Each MC question was converted to three TF questions (two false TF statements and one true TF statement). To ensure that corresponding MC and TF items were as comparable as

possible in reading time and in other aspects, the stem was included in each option of the MC items as it necessarily was in each TF item. The conversion process is illustrated by the following example:

MC item: Circle the letter corresponding to the best statement for each item.

a. The mode may have more than one value in the same distribution.
b. The range may have more than one value in the same distribution.
c. The standard deviation may have more than one value in the same distribution.

TF item: Circle "T" if the statement is true and "F" if the statement is false.

1. T F The mode may have more than one value in the same distribution.
2. T F The range may have more than one value in the same distribution.
3. T F The standard deviation may have more than one value in the same distribution.

Each test was then divided into three equal parts. Each of the three TF items corresponding to a MC item was placed on a separate part. This was done in an attempt to minimize dependency between response to a TF item and response to a similar previous item. Items on each part were then randomly ordered. All students were instructed to complete part one and hand it in, then to pick up and complete the second and then the third part. Items stressed application and interpretation of concepts rather than memorization of facts. The MC midterm consisted of 40 questions: 14, 13 and 13 items on each part; the TF midterm had 120 questions, 40 items on each part. Test forms were randomly ordered and distributed to students.

The class was given 90 minutes to complete the exam. Since all subjects finished within this time speed was not considered to be a factor influencing performance. Subjects were interrupted after 12 minutes and asked to circle the number of the item on which they were working. These data were used to determine the number of TF items answered per MC item.

The students' cumulative grade points were used as an external criterion from which a concurrent validity coefficient was calculated.

RESULTS

A Kuder-Richardson formula 20 reliability coefficient was computed for each of the two test forms. The reliability of the MC test was then adjusted with the Spearman-Brown formula to equate testing time. Since subjects responded to 1.19 MC items and 2.85 TF items per minute, the TF test took 1.25 times as long to complete as the MC test. The reliability estimate of the MC form was adjusted by this factor. Means, standard deviations, reliability coefficients, and the correlation between each test form and GPA are presented in Table 1. A statistical test of the hypothesis that reliability coefficients associated with two different measurement procedures are equal has been developed and empirically examined by Feldt (1969). The statistic is based on the assumption that the scores on k parallel parts of a test instrument conform to the assumptions of the two-factor random model of analysis of variance: a normally distributed population randomly sampled and homogeneity of variance for the k parts of the test. The difference between the reliabilities of the MC and TF forms was tested using this

statistic and was not found to be significant (W=1.35 with 23 and 24 degrees of freedom).

To compare the validities of the two forms, a Pearson product moment correlation was calculated between test score and GPA for each test form (Table 1). The difference was tested using a Fisher's Z transformation, the obtained value of z=.13 not reaching significance

Table 1

Item Format	$\bar{x}$	s	N	# of items	Unadjusted KR20	(w)	Adjusted KR20	(w)	$r_{x \cdot GPA}$	(z
TF	85.80	9.80	25	120	.80		(.80)		.49	
						(1.6)		(1.35)		(.
MC	29.12	4.30	25	40	.68		.73		.52	

DISCUSSION

The ratio of TF to MC items responded to per unit time in this study (2.4) differs from those reported by Oosterhof and Glasnapp (1.73) and Frisbie (1.5) (1974). This finding supports Frisbie's (1973) statement that different student groups and different examination topics may produce variant response patterns. Since three TF items are theoretically equivalent to one MC item, one would posit that the ratio of TF:MC items answered per unit time should approach the ratio of number of MC options:1 if reading time were constant. Models of the optimal number of choices per item have assumed total testing time to be proportional to the number of choices per item but empirical studies including those referenced above have shown that TF and MC item types do not satisfy this assumption.

The TF:MC ratio found in this study is higher than that found in previous studies. This suggests that students spent comparatively less time per TF item or comparatively more time per MC item. The format of the MC items differs from that standardly used and may have slowed students' processing of the MC items.

Results indicate that reliabilities of the TF forms can be as high as reliabilities of three-option MC forms and can be as effective in measuring classroom achievement. These results contradict other findings regarding the reliability of the TF form. This difference may be due in part to the use in this study of a TF test which was longer than the MC test and allowed a better estimate of reliability for this form. The smaller number of items on the MC form was likely to be a possible factor in its lower obtained reliability. Scores were not corrected for guessing and the range of the TF form was not restricted as it had been in previous studies. Rather, the range of the MC form was comparatively restricted. Differing also was the method of comparing formats, the present study employing all options of corresponding MC items as TF items with no known initial biases in item discrimination. Another factor favoring heightened reliability of the TF form was the ratio (.67) of items keyed false to those keyed true. Frisbie (1974) suggested that false items generally discriminate better than true items, 60% false being suggested as a possible optimum (Ebel, 1972).

Results of this study provide support for the use of TF teacher-made tests as alternatives to MC tests with no loss in reliability or validity. However, as previous studies have shown, these results are not obtained when MC items have been subjected to revision and

the range of the TF form is restricted. It is suggested that a further comparison of formats be made in which both types of items have been improved and matched for difficulty and discrimination levels. Also, further investigation is suggested varying the number of TF items used as MC options and varying the ratio of false to true TF statements.

BIBLIOGRAPHY

Blood, D.F. & Budd, W.S. Educational measurement and evaluation. New York: Harper & Row, Publishers, 1972.

Ebel, R.L. Essentials of educational measurement. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1972.

Feldt, L.S. A test of the hypothesis that Cronbach's Alpha or Kuder Richardson coefficient twenty is the same for two tests. Psychometrika, 1969, 34, 363-373.

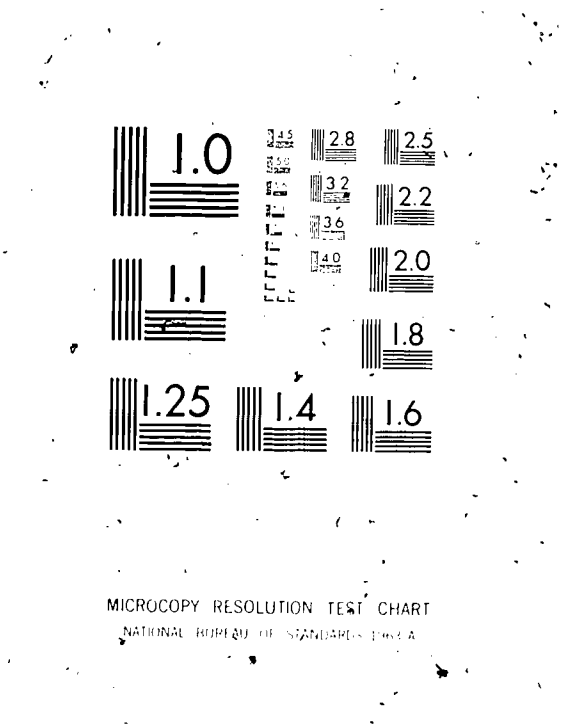
Frisbie, D.A. Multiple choice versus true-false: a comparison of reliabilities and concurrent validities. Journal of Educational Measurement, 1973, 10, 297-304.

Frisbie, D.A. The effect of item format on reliability and validity: a study of multiple choice and true-false achievement tests. Educational and Psychological Measurement, 1974, 34, 885-892.

Karmel, L.J. Measurement and evaluation in the schools. London: Macmillan Co., 1970.

Nunnally, J.C. Educational measurement and evaluation. New York: McGraw-Hill Book Co., 1964.

Oosterhof, A.C. & Glasnapp, D.R. Comparative reliabilities and difficulties of the multiple-choice and true-false formats. Journal of Experimental Education, 1974, 42, 62-64.


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963 A

DOCUMENT RESUME

ED 169 073 TM 007 865

AUTHOR Green, Kathy
TITLE Multiple Choice Converted to True-False: Comparative Reliabilities and Validities.
PUB DATE 78
NOTE 10p.; Paper presented at the Annual Meeting of the Western Psychological Association (San Francisco, California, 1978)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Tests; *Comparative Testing; Higher Education; *Multiple Choice Tests; Reaction Time; Response Style (Tests); *Test Construction; *Test Reliability; *Test Validity
IDENTIFIERS Teacher Developed Tests; *True False Tests

ABSTRACT

Forty three-option multiple choice (MC) statements on a midterm examination were converted to 120 true-false (TF) statements, identical in content. Test forms (MC and TF) were randomly administered to 50 undergraduates, to investigate the validity and internal consistency reliability of the two forms. A Kuder-Richardson formula 20 reliability was computed for each form. Reliability of the MC form was then adjusted with the Spearman-Brown formula to equate testing time, since the MC form took three-fourths as much time to complete as the TF form. Adjusted reliability coefficients of the TF and MC forms were .80 and .73, respectively. To compare validity, a Pearson product moment correlation was computed between test score and grade point average; validity coefficients were .49 (TF) and .52 (MC). Results support the use of TF teacher made tests as alternatives to MC tests with no loss in reliability or validity. However, as previous studies have shown, these results are not obtained when MC items are revised and the range of the TF form is restricted. (CP)

MULTIPLE CHOICE CONVERTED TO TRUE-FALSE: COMPARATIVE RELIABILITIES AND VALIDITIES

Kathy Green

Paper presented at the Western Psychological Association Annual Convention, 1978, San Francisco.

The advantages and disadvantages of multiple-choice and true-false formats have been studied by a number of investigators (e.g. Nunnally, 1964; Karmel, 1970; Blood & Budd, 1972). One advantage claimed for the true-false (TF) test is that it allows a more efficient sampling of course content. Proponents of the multiple-choice (MC) format, however, argue that this advantage may be offset by lowered reliability coefficients that occur primarily as the result of guessing on TF items. Evidence is available to support the contention that reliability (Oosterhof & Glasnapp, 1974) if not validity (Frisbie, 1973, 1974) of the MC test form is higher than that of the TF test form.

Empirical comparisons of the TF and MC formats, however, have been beset by a number of methodological problems. One serious difficulty is that greater care is often given to the preparation of MC items than to TF items. For example, more MC than TF tests have been item analyzed and revised prior to their being compared (Frisbie, 1973, 1974; Oosterhof & Glasnapp, 1974). It would be expected that if MC items were more extensively revised than TF items, the reliability of the MC form would be higher. The present study employed teacher-made tests in which the MC and TF items were not differentially improved.

To further complicate matters, TF items are often constructed from parallel MC items on either a one-to-one or on a two-to-one basis--either one or two questions being generated from each MC item. Resultant test forms then have contained as many or twice as many TF as MC items. However, even at a ratio of 2:1 TF:MC questions, an estimate of the hypothetically lengthened TF form reliability was necessary to equate testing time (Oosterhof & Glasnapp, 1974).

Frisbie (1973) suggested that use of a longer TF test would probably increase the variance of the TF scores and produce a better estimate of the relationship between MC and TF forms. The present study converted each three-option MC statement to three separate parallel TF statements, obviating the need to adjust the reliability of the TF form to equate testing time and providing for increased variance of the TF scores.

The expectations for this study were:

1. When MC items are converted to TF items, the internal consistency reliabilities of the two forms do not differ significantly.
2. When testing time is equated, reliabilities of the two forms do not differ significantly.
3. Validity of the two forms does not differ significantly.

To reduce likelihood of Type II errors, null hypotheses required rejection at the .10 rather than the .05 level.

METHOD

Subjects

This study was conducted during the summer quarter of 1977. Fifty undergraduates enrolled in a required introductory class in tests and measurement at the University of Washington served as subjects. Subjects were naive at the time of testing regarding the nature of reliability and the relationship between MC and TF item formats.

Instrumentation

A MC and TF form of a midterm examination were constructed for the class. Items on the two forms differed in format only, the content being identical in every instance. Each MC question was converted to three TF questions (two false TF statements and one true TF statement). To ensure that corresponding MC and TF items were as comparable as

possible in reading time and in other aspects, the stem was included in each option of the MC items as it necessarily was in each TF item. The conversion process is illustrated by the following example:

MC item: Circle the letter corresponding to the best statement for each item.

a. The mode may have more than one value in the same distribution.
b. The range may have more than one value in the same distribution.
c. The standard deviation may have more than one value in the same distribution.

TF item: Circle "T" if the statement is true and "F" if the statement is false.

1. T F The mode may have more than one value in the same distribution.
2. T F The range may have more than one value in the same distribution.
3. T F The standard deviation may have more than one value in the same distribution.

Each test was then divided into three equal parts. Each of the three TF items corresponding to a MC item was placed on a separate part. This was done in an attempt to minimize dependency between response to a TF item and response to a similar previous item. Items on each part were then randomly ordered. All students were instructed to complete part one and hand it in, then to pick up and complete the second and then the third part. Items stressed application and interpretation of concepts rather than memorization of facts. The MC midterm consisted of 40 questions: 14, 13 and 13 items on each part; the TF midterm had 120 questions, 40 items on each part. Test forms were randomly ordered and distributed to students.

The class was given 90 minutes to complete the exam. Since all subjects finished within this time speed was not considered to be a factor influencing performance. Subjects were interrupted after 12 minutes and asked to circle the number of the item on which they were working. These data were used to determine the number of TF items answered per MC item.

The students' cumulative grade points were used as an external criterion from which a concurrent validity coefficient was calculated.

RESULTS

A Kuder-Richardson formula 20 reliability coefficient was computed for each of the two test forms. The reliability of the MC test was then adjusted with the Spearman-Brown formula to equate testing time. Since subjects responded to 1.19 MC items and 2.85 TF items per minute, the TF test took 1.25 times as long to complete as the MC test. The reliability estimate of the MC form was adjusted by this factor. Means, standard deviations, reliability coefficients, and the correlation between each test form and GPA are presented in Table 1. A statistical test of the hypothesis that reliability coefficients associated with two different measurement procedures are equal has been developed and empirically examined by Feldt (1969). The statistic is based on the assumption that the scores on k parallel parts of a test instrument conform to the assumptions of the two-factor random model of analysis of variance: a normally distributed population randomly sampled and homogeneity of variance for the k parts of the test. The difference between the reliabilities of the MC and TF forms was tested using this

statistic and was not found to be significant (W=1.35 with 23 and 24 degrees of freedom).

To compare the validities of the two forms, a Pearson product moment correlation was calculated between test score and GPA for each test form (Table 1). The difference was tested using a Fisher's Z transformation, the obtained value of z=.13 not reaching significance.

Table 1

Item Format	$\bar{x}$	s	N	# of items	Unadjusted KR20	(w)	Adjusted KR20	(w)	$r_{x \cdot GPA}$	(z)
TF	85.80	9.80	25	120	.80		(.80)		.49	
						(1.6)		(1.35)		(.13)
MC	29.12	4.30	25	40	.68		.73		.52	

DISCUSSION

The ratio of TF to MC items responded to per unit time in this study (2.4) differs from those reported by Oosterhof and Glasnapp (1.73) and Frisbie (1.5) (1974). This finding supports Frisbie's (1973) statement that different student groups and different examination topics may produce variant response patterns. Since three TF items are theoretically equivalent to one MC item, one would posit that the ratio of TF:MC items answered per unit time should approach the ratio of number of MC options:1 if reading time were constant. Models of the optimal number of choices per item have assumed total testing time to be proportional to the number of choices per item but empirical studies including those referenced above have shown that TF and MC item types do not satisfy this assumption.

The TF:MC ratio found in this study is higher than that found in previous studies. This suggests that students spent comparatively less time per TF item or comparatively more time per MC item. The format of the MC items differs from that standardly used and may have slowed students' processing of the MC items.

Results indicate that reliabilities of the TF forms can be as high as reliabilities of three-option MC forms and can be as effective in measuring classroom achievement. These results contradict other findings regarding the reliability of the TF form. This difference may be due in part to the use in this study of a TF test which was longer than the MC test and allowed a better estimate of reliability for this form. The smaller number of items on the MC form was likely to be a possible factor in its lower obtained reliability. Scores were not corrected for guessing and the range of the TF form was not restricted as it had been in previous studies. Rather, the range of the MC form was comparatively restricted. Differing also was the method of comparing formats, the present study employing all options of corresponding MC items as TF items with no known initial biases in item discrimination. Another factor favoring heightened reliability of the TF form was the ratio (.67) of items keyed false to those keyed true. Frisbie (1974) suggested that false items generally discriminate better than true items, 60% false being suggested as a possible optimum (Ebel, 1972).

Results of this study provide support for the use of TF teacher-made tests as alternatives to MC tests with no loss in reliability or validity. However, as previous studies have shown, these results are not obtained when MC items have been subjected to revision and

the range of the TF form is restricted. It is suggested that a further comparison of formats be made in which both types of items have been improved and matched for difficulty and discrimination levels. Also, further investigation is suggested varying the number of TF items used as MC options and varying the ratio of false to true TF statements.

BIBLIOGRAPHY

Blood, D.F. & Budd, W.S. Educational measurement and evaluation. New York: Harper & Row, Publishers, 1972.

Ebel, R.L. Essentials of educational measurement. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1972.

Feldt, L.S. A test of the hypothesis that Cronbach's Alpha or Kuder Richardson coefficient twenty is the same for two tests. Psychometrika, 1969, 34, 363-373.

Frisbie, D.A. Multiple choice versus true-false: a comparison of reliabilities and concurrent validities. Journal of Educational Measurement, 1973, 10, 297-304.

Frisbie, D.A. The effect of item format on reliability and validity: a study of multiple choice and true-false achievement tests. Educational and Psychological Measurement, 1974, 34, 885-892.

Karmel, L.J. Measurement and evaluation in the schools. London: Macmillan Co., 1970.

Nunnally, J.C. Educational measurement and evaluation. New York: McGraw-Hill Book Co., 1964.

Oosterhof, A.C. & Glasnapp, D.R. Comparative reliabilities and difficulties of the multiple-choice and true-false formats. Journal of Experimental Education, 1974, 42, 62-64.