DOCUMENT RESUME

ED 129 916   TM 005 790

AUTHOR   Traub, Ross E.; Fisher, Charles W.
TITLE   On the Equivalence of Constructed-Response and
               Multiple-Choice Tests.
SPONS AGENCY   Ontario Inst. for Studies in Education, Toronto.
NOTE   31p.

EDRS PRICE   MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS   Memory; *Multiple Choice Tests; Recall
               (Psychological); *Response Mode; Risk
IDENTIFIERS   *Constructed Response Item Format; *Coombs Multiple
               Choice Item Format; Following Directions; Test
               Items

ABSTRACT
               Two sets of mathematical reasoning and two sets of
verbal comprehension items were cast into each of three
formats--constructed response, standard multiple-choice, and Coombs
multiple-choice--in order to assess whether tests with indentical
content but different formats measure the same attribute, except for
possible differences in error variance and scaling factors. The
resulting 12 tests were administered to 199 eighth-grade students.
The hypothesis of equivalent measures was rejected for only two
comparisons: the constructed response measure of verbal comprehension
was different from both the standard and the Coombs multiple-choice
measures of this ability. Maximum likelihood factors analysis
confirmed the hypothesis that a five factor structure will give a
satisfactory account of the common variance among the 12 tests. As
expected, the two major factors were mathematical reasoning and
verbal comprehension. Contrary to expectation, only one of the other
three factors bore a (weak) resemblance to a format factor. Tests
marking the ability to follow directions, recall and recognition
memory, and risk taking were included, but these variables did not
correlate as expected with the three minor factors. (Author/MV)

On the Equivalence of Constructed-Response and

Multiple-Choice Tests[1, 2]

Ross E. Traub

Ontario Institute for Studies in Education

and

Charles W. Fisher

Far West Laboratory for Educational Research and Development

2

On the Equivalence of Constructed-Response and

Multiple-Choice Tests

Ross E. Traub

Ontario Institute for Studies in Education

Charles W. Fisher

Far West Laboratory for Educational Research and Development

Abstract

Two sets of mathematical reasoning and two sets of verbal comprehension
items were cast into each of three formats--constructed response, standard
multiple-choice, and Coombs multiple-choice--in order to assess whether
tests with identical content but different formats measure the same
attribute, except for possible differences in error variance and scaling
factors. The resulting 12 tests were administered to 199 eighth-grade
students. The hypothesis of equivalent measures was rejected for only two
comparisons: the constructed response measure of verbal comprehension was
different from both the standard and the Coombs multiple-choice measures of
this ability. Maximum likelihood factor analysis confirmed the hypothesis
that a five factor structure will give a satisfactory account of the common
variance among the 12 tests. As expected, the two major factors were
mathematical reasoning and verbal comprehension. Contrary to expectation,
only one of the other three factors bore a (weak) resemblance to a format
factor. Tests marking the ability to follow directions, recall and recogni-
tion memory, and risk taking were included, but these variables did not
correlate as expected with the three minor factors.

3

A question of enduring interest for students of educational measurement is whether tests that employ different response formats, but that in other respects are as similar as possible, measure the same attribute. This question has been asked of constructed-response as compared with multiple-choice tests (Cook, 1955; Davis & Fifer, 1959; Heim & Watts, 1967; Vernon, 1962) and of multiple-choice tests having standard as compared with non-standard formats (Dressel & Schmid, 1953; Coombs, Milholland & Womer, 1956; Rippey, 1968; Hambleton, Roberts & Traub, 1970). Results of available research suggest that the distributions of scores on tests employing different formats cannot be assumed to have the same mean and standard deviation, even when the tests are administered to the same group of examinees or to groups that differ only because of random allocation of examinees to groups. In addition, the reliability and criterion correlation coefficients associated with different response formats cannot be assumed to be the same. These results are not, of course, sufficient evidence to reject the null hypothesis that tests with different formats measure the same attribute. It is possible to account for differences in means and standard deviations through appeal to possible differences in the scales of measurement associated with different formats; and differences in reliability and criterion correlation coefficients can be attributed to possible differences in the relative amount of error variance associated with the different test formats and also to possible differences in the scales of measurement such that one scale is a nonlinear transformation of the other.

4

For several years now statistical procedures have been in existence for testing the null hypothesis of equivalence of measures. Early exemplars of these procedures (Lord, 1957; McNemar, 1958) were somewhat difficult to use. More recently, however, Lord (1971) has presented a statistically rigorous test based on work by Villegas (1964) that is relatively easy to employ. This test is of the hypothesis that "two sets of measurements differ only because of a) errors of measurement, b) differing units of measurement, and c) differing arbitrary origins for measurement" (Lord, 1971, p.1). Clearly, this test accounts for all the previously described reasons for differences between the measurements yielded by two different test formats except those differences caused by the fact that one scale is a nonlinear transformation of the other.

In addition to Lord's procedure, recent developments in factor analysis make it possible to test hypotheses about the relationship among measurements arising from tests with different formats. These developments, subsumed under the heading, confirmatory factor analysis, have been made principally by Jöreskog (1969, 1971) and McDonald (1969).

The purpose of the present investigation was to test the equivalence of three response formats, each applied to items from two different content domains. The formats were (i) constructed-response, (ii) standard multiple-choice, in which the examinee is instructed to choose one option per item, the one he thinks is correct, and (iii) non-standard multiple-

5

choice, in which the examinee is asked to identify as many of the incorrect

options as he can. This latter procedure was described by Coombs,

Milholland and Womer (1956) and is hereafter called the Coombs format or

the Coombs procedure. The two content domains were verbal comprehension,

as defined operationally by questions on the meaning of words, and mathe-

matical reasoning, as defined operationally by questions about a variety

of mathematical concepts and skills, and by problems, the solution of which

depends on the ability to apply a variety of mathematical concepts and

skills.

The motivation for studying the equivalence of measurements arising

from different response formats was to gain some further understanding of

partial knowledge. The standard multiple-choice format does not assess and

credit partial knowledge, the kind of knowledge that enables an examinee to

respond at a better-than-chance level to items that cannot with certainty

be answered correctly. The Coombs format nullifies this criticism because

it enables an examinee to gain partial credit by identifying one or more

of the incorrect options to an item, even when not all of the incorrect

options can be identified. What remains at issue, in the face of this

logical analysis, is whether measurements based on the Coombs format

reflect the same attribute as measurements based on the standard multiple-

choice format. For example, it might be the case that the longer and more

involved instructions associated with the Coombs format introduce the

factor of following directions[3] into the measurements, a factor that might

not be present in measurements based on the standard multiple-choice for-

mat with its simpler instructions.

6

A comparison of the Coombs and standard multiple-choice formats appears interesting in its own right, but both these formats can be viewed as ways to simplify and objectify the scoring of constructed-response items. To the extent that this view of objectively scorable tests is accepted, interest extends to a comparison of measurements derived from all three formats. Again, the issue is whether the measurements derived from a constructed-response format reflect the same attribute as measurements derived from objectively scorable formats. For example, items that are designed to test factual knowledge and that involve the constructed-response format can be answered by the exercise of recall memory. The same items when cast into a multiple-choice format can be answered by the exercise of either recall or recognition memory. In addition, multiple-choice formats are more clearly subject to the influence of risk taking (guessing) behavior than is a constructed-response format. In the case of the constructed-response format, an examinee can guess only if he makes the effort to generate a response. This fact alone operates against risk-taking behavior. In addition, the set of possible responses is probably quite large, although for any examinee it consists of only those possibilities he can generate and this number is not necessarily large; the larger the set of possible responses, the less likely the examinee is to guess correctly and the less that risk-taking can influence his test score. On the other hand, in the case of choice formats, the set of possible responses is small, in addition to being precisely the same for every examinee. This means that the probability of a correct guess is sufficiently large for risk-taking to influence test scores significantly.

7

Fortunately, the topic of risk-taking on multiple-choice tests has been the subject for considerable research and measures of individual differences in risk-taking have been proposed; hence it is a factor that can be included as an independent variable in research studies.

In summary, the main purpose of the present study was to test the equivalence of measurements obtained using constructed-response, standard multiple-choice and Coombs response formats. Secondarily, the study was designed to identify format factors and to study the association between these factors, if found, and the psychological attributes of following directions ability, recall memory, recognition memory, and risk taking.

Method

Instrumentation

To attain the main purpose of this investigation, it was necessary to impose two constraints on the measures devised for each content domain: (1) The content of the measures for one test format had to be as similar as possible to the content of the measures for another test format. This constraint was satisfied by using the same set of item stems for all three test formats and the same item response options for both the standard multiple-choice and Coombs formats. (2) The number of measures per response format had to be at least two in order to implement Lord's (1971) procedure for testing equivalence. This constraint was satisfied by forming two sets of verbal comprehension items and two sets of mathematical reasoning items. The two sets of verbal comprehension items were drawn from a pool formed by the items marking the verbal comprehension factor in the Kit of Reference Tests for Cognitive Factors (French, Ekstrom & Price, 1963); the two sets of mathematical reasoning items were drawn from a pool consisting of the mathematics items in Forms 3A and 4A of the 1957 edition of SCAT (ETS, 1957),[4]

the items marking the general reasoning factor in the Kit of Reference Tests for Cognitive Factors (French, et al, 1963), and the items in the Canadian New Achievement Test in Mathematics (OISE, 1965). The large pools of verbal comprehension and mathematical reasoning items were pretested in their standard multiple-choice formats, and under instructions to answer every item with no penalty for wrong answers, to approximately 100 students at the same eighth-grade level as the students who subsequently participated in the study proper. These pretest data were used to compute indices of item difficulty--the percentage of correct responses--and item discrimination-- the item-total biserial correlation coefficient. (The total score used in the computation of a biserial correlation coefficient was the sum of scores on all the items included in the pool for a given content domain.) The two sets of items drawn from the verbal comprehension pool each contained 50 items, the two sets of items from the mathematical reasoning pool each contained 30 items. The item sets for a content domain were matched for pretest indices of difficulty, with the average difficulty being .50 in each case. The item sets were also matched as closely as possible for values of the pretest indices of discrimination.

The secondary purpose of the study was to seek response format factors and, if such factors were isolated, to study the degree of association between the factors and measures of possibly related psychological attributes. The search for format factors, given the design of the study, took place among the covariances between measures having the same format but different content. In other words, a factor defined by the constructed-response format was conceived as one that

9

would be associated with the constructed-response measures of both the verbal comprehension and mathematical reasoning domains of content and not with the standard multiple-choice and Coombs measures of these domains. Format factors associated with the standard multiple-choice and Coombs formats would be similarly defined.

The variables of following directions, recall memory, recognition memory and propensity for risk taking (on multiple-choice tests) were measured for the purpose of studying the association between these variables and format factors, if such factors were identified. The ability to follow directions was measured by two instruments that had been used previously by Traub (1970) and prepared as adaptations of a test devised originally by J. W. French. Two measures of recall memory were employed, both of which were taken from the tests of associative memory contained in the <u>Kit of Reference Tests for Cognitive Factors</u> (French, Ekstrom & Price, 1963). Recognition memory was assessed by two measures, both adaptations of materials developed by Duncanson (1966).[5] The fourth variable, risk taking, was measured using an instrument developed by Traub and Hambleton (1972). The rationale for this instrument was proposed by Swineford (1938; 1941).

<u>Design and Subjects</u>

Lord's procedure for testing the equivalence of measures and the method of confirmatory factor analysis are applicable to data obtained from a single group of subjects. In this study usable data were obtained on 199 eighth grade students (93 females), with a mean age on September 1, 1971 of approximately 13 years, 8 months (the standard deviation of the age distribution was approximately 8 months). During the 1971-72 academic

year these students attended one of the two junior high schools that

cooperated in the study. Both cooperating schools were located in East

York, a borough of Metropolitan Toronto. The neighborhoods served by

these schools were described by the school principals as a mixture of the

lower and middle classes.[6]

In any study involving tests that differ only in response format,

care must be taken to minimize memory effects. This was done in the

present study by scheduling the tests so that there was a two-week interval

between each administration of the same set of items and by administering

the constructed response formats first--Heim and Watts (1966) found that

carry-over from one administration of a set of vocabulary items to a

second administration of the items in a different format was markedly less

when the constructed-response format preceded the multiple-choice format

than when the reverse order was followed.

It was anticipated, and subsequent events tended to confirm that

motivating students to work all versions of the verbal comprehention and

mathematical reasoning tests would be a problem. The following steps were

taken to minimize this problem: (i) students were told at the first

administration that the study was designed to find out whether people

score better when a test involves one kind of response format than when it

involves another, that they would be tested periodically over a period of

weeks and that their scores on the tests would be sent to them individually

(this promise was kept in that copies of individual report forms were

delivered to the school when the scoring had been completed); (ii) the

standard multiple-choice format was introduced with the comment that it

would give the student a chance to improve his performance on the

the constructed-response tests; (iii) the Coombs format was introduced as another chance to improve on past performance.

Two other critical conditions of the test administrations were the scoring instructions and the time limits provided for the administration of each test. On all tests employing a constructed-response format--two verbal comprehension, two mathematical reasoning, two following directions and two recall memory tests--students were informed that the number of correct answers would be the score on these tests and that, on the verbal comprehension and mathematical reasoning tests with a constructed-response format, it was to their benefit to show all their work because partial credit could be obtained for work done on questions answered incorrectly. Six of the remaining measures, four verbal comprehension and mathematical reasoning tests and two tests of recognition memory, were presented in a multiple-choice format and were scored for the number of correct answers; in view of this, the students were instructed to answer every question,and guess if necessary. The four tests presented in the-Coombs format and the measure of risk taking involved rather elaborate scoring instructions with a complex system of rewards and penalties. The students were informed of the scoring system in each case and several examples were considered to demonstrate the potential effect of the scoring system.

As regards the time limits for the tests, they are reported in Table 1 for each test. These limits were established on the basis of

_____

Insert Table 1 about here

_____

pilot administrations of the tests and (except in the case of the memory
tests) were generous, even for the Coombs format which was most time con-
suming, so as to achieve essentially power conditions. The time limits for
the tests of recall memory are those specified by French et al (1963);
the limits for the recognition memory tests were set on the basis of pre-
test results to achieve a satisfactory distribution of scores (i.e. a
distribution with a range approximately three standard deviations either
side of the mean).

Scoring

Special keys were prepared for the constructed-response versions
of the verbal comprehension and mathematical reasoning tests. These keys,
which indicated to the marker how to award partial credit for certain wrong
answers were applied to responses obtained from a pretest and were revised
as required in the light of apparent inadequacies. The final forms of the
keys were applied by independent scorers to a random sample of 50
constructed-response answer booklets from one school for Form B of the tests
for both content domains. The correlation between the marks assigned by
the scorers was 0.97 for the verbal comprehension test and 0.98 for the
mathematical reasoning test.

All other tests used in the study could be scored objectively.

(Copies of tests and scoring keys are available from the authors
on request.)

A Note on Sample Size

The sample size of 199 represents approximately one-half the total
number of eighth-grade students attending the two cooperating schools
during the time the data were being collected. The data from the other

students were discarded for one of several reasons: (i) some students were so-called New Canadians and had difficulty understanding written English; (ii) some students were absent from school on one or more of the seven days on which the tests were administered; (iii) some students attempted fewer than ten of the questions on a test or marked their multiple-choice answer sheets following a clear pattern unrelated to the pattern of correct answers and were judged to have paid little attention to the task; (iv) some students were observed to copy answers from other students during one or more of the testing occasions. The frequency of occurrence of reasons (iii) and (iv) was zero for the first two testing occasions but over the next four occasions, when the test items were repeated in the different formats, this frequency departed quite substantially from zero. The occurrence of this type of behavior indicates the difficulty that is encountered in sustaining student motivation when tests are administered repeatedly.

Results and Discussion

Basic Statistics

Means and standard deviations for all 19 measures, coefficients α for the 12 mathematical reasoning and verbal comprehension measures, and intercorrelations amongst all 19 measures are presented in Table 1.

Alpha coefficients are not reported for the seven marker variables; the calculation of α is impossible for the measure of risk-taking and cannot be justified for speeded tests such as the tests of recall and recognition memory. Despite this, the results suggest that the reliabilities of at least the memory tests were relatively low. The evidence for this suggestion consists of the correlations between the pairs of tests

designed to measure the recall and recognition memory factors. Although these pairs of tests are not parallel in content, their intercorrelations are much lower than would be expected for tests that reliably measure the same ability.

Equivalence of Measures

All possible pairs of tests having the same content and different formats were assessed for equivalence using Lord's (1971) procedure.[7,8] On the basis of the results achieved for the measures of mathematical reasoning, the hypothesis of equivalence cannot be rejected for any of the three possible contrasts of test formats. The results for the measures of verbal comprehension indicate that the hypothesis of equivalence can be rejected for two contrasts--constructed-response vs. standard multiple-choice and constructed-response vs. Coombs. It was not possible to reject the null hypothesis of equivalence for the contrast between the standard multiple-choice and Coombs formats.

To ascertain whether factors associated with test format would override those associated with content, three other pairings were considered. In each case response format was held constant and content was varied; that is, the constructed response versions of mathematical reasoning and verbal comprehension were tested for equivalence, as were the standard multiple-choice and Coombs versions of these tests. The hypothesis of equivalence was rejected for all three of these comparions.

The foreging results indicate that the tests of mathematical reasoning measured the same attribute regardless of response format, whereas the attributes measured by tests of verbal comprehension varied as a function of response format. A conception of mental funtioning that would

15

account for this finding is the following. (1) Determining the correct answer to a mathematical reasoning item involves working out the answer to the item regardless of the format of the test. The work is recorded in the case of constructed response items and is used as a basis for choosing a response in the case of the standard multiple-choice and Coombs formats. (2) Determining the correct answer to a verbal comprehension item involves recalling definitions when a constructed-response format is used. When standard multiple-choice and Coombs formats are used, however, it is only necessary to recognize the correct answer, and recognition is facilitated by ruling out implausible response options. This conception of the difference in mental operations that are employed in working mathematical reasoning as compared with verbal comprehension tests carries the implication that the main advantage of the Coombs response format--the possibility of revealing partial knowledge by identifying one or more response options as incorrect but not identifying all the incorrect options--would be utilized to a greater extent with the verbal comprehension than the mathematical reasoning items. Statistics confirming this implication are reported in Table 2.[9]

Insert Table 2 about here

Format Factors

The intercorrelations among the 12 measures of mathematical reasoning and verbal comprehension were subjected to confirmatory factor analysis using the COSA-I program of McDonald and Leong (Note 1) in an effort to identify format factors. A format factor is by definition a factor associated with tests employing the same response format, regardless of test content. In line with the hypothesized existence of format factors is a five factor structure involving two correlated factors, one marking

16

mathematical reasoning, the other marking verbal comprehension, and three

orthogonal format factors, one for each of the three response formats

included in the study. It proved possible to obtain a satisfactory fit of

a five-factor structure provided that, in addition to the specified five

factors, the possibility of correlated unique factors among all six measures

of mathematical reasoning and among all six measures of verbal comprehension

was allowed. For this structure, the approximate χ^2 statistic arising from

the goodness-of-fit test was 21.297, which, with 11 degrees of freedom, has

a probability of chance-occurrence under the null hypothesis of slightly more

than 0.03. Estimated values of the parameters of this structure are given

in Table 3.

Insert Table 3 about here

Any attempt to deviate from this structure by fitting fewer factors

resulted in significantly larger χ^2 statistics--the increase in the value of

χ^2 was approximately equal to twice the increase in the number of degrees of

freedom associated with each decrement of one in the number of factors from

five to two factors. Any attempt to equate to zero some or all of the corre-

lations among unique factors for tests having the same content resulted

in the occurrence of a so-called Heywood case, in which the estimated unique

variance of at least one of the 12 tests was a non-trivial but meaningless

negative number.

There are several points worth noting about the structure reported

in Table 3:

(i) The two main factors are mathematical reasoning (Factor I)

and verbal comprehension (Factor II). As expected these factors are highly

intercorrelated.

(ii) The sets of items comprising forms A and B of each content domain are far from parallel, regardless of the format in which they are presented. Had these item sets been parallel, the factor coefficients and unique variances of forms A and B for a given format and content domain would have been the same. But when this kind of constraint was imposed on the structure fitted to the data, the result was very unsatisfactory.

(iii) The fitted structure ignores a substantial amount of the variance held in common among the six tests of verbal comprehension. This is clear from the size of the off-diagonal entries in U for the six verbal tests. These entries are considerably larger on average than the corresponding entries for the six mathematical reasoning tests.

(iv) It was hoped that factors three, four and five would look like format factors. In order to have the appearance of a format factor, the coefficients of the four tests sharing the same response format should all have the same algebraic sign and be large enough in absolute magnitude to be distinguishable from zero. The only factor of the three that comes close to satisfying these conditions is the third, which can, perhaps, be called a constructed-response factor. The coefficients on the fourth and fifth factors, however, do not meet the conditions for format factors.

As a further guide to interpreting the factor structure reported in Table 3, an "extension analysis" (Lord, 1956, pp. 40, 42) was performed in which least squares estimates of the coefficients of the seven marker variables on the five factors were obtained. These coefficients are reported in Table 4. Several observations are supported by the

---

Insert Table 4 about here

---

numbers given in the table:

(i) The tests of following directions ability have sizeable coefficients on the mathematical reasoning and the verbal comprehension factors (I and II, respectively). These tests do not, contrary to expectation, have substantially larger coefficients on the fifth factor, the one defined by tests with the Coombs format, than they have on the third and fourth factors, those defined by the constructed-response and standard multiple-choice tests, respectively.

(ii) The results for the tests of recall memory are interesting in that they have positive coefficients on the mathematical reasoning factor and negative coefficients on the verbal comprehension factor. The positive coefficients on mathematical reasoning may reflect nothing more than that the two recall memory tests required examinees to form associations between pictures or object labels and numbers. It is possible, however, to use these results as partial support for the previously described theory of examinee behavior on constructed-response as compared with multiple-choice tests. According to the theory, examinees respond to mathematical reasoning items, regardless of test format, by doing the operations needed to derive answers to the questions. This is an activity which presumably would draw heavily on recall memory. The factor structure provides support for this suggestion. The theory also predicts (a) a positive association between recall memory and constructed-response tests of verbal comprehension and (b) a positive association between recognition memory and multiple-choice tests of verbal comprehension. Because the verbal comprehension factor in this study is marked by both constructed response and multiple-choice tests, it is difficult to predict just what associations there should be between the verbal comprehension factor and the tests of recall memory and recognition memory. The obtained negative coefficients for recall memory

on the verbal comprehension factor are something of a puzzle--why should performance of these tests be hampered by recall memory?--but the positive coefficients for recognition memory on this factor are not surprising, although their size is smaller than might be expected.

(iii) Neither the recall memory nor the recognition memory tests had coefficients the size they were expected to have on the factors marked by tests with different formats, i.e., high coefficients for recall memory on the third factor and high coefficients for recognition memory on the fourth and fifth factors.

(iv) The positive coefficient for the measure of risk taking on the verbal comprehension factor is most probably a reflection of the fact that the content of the risk-taking measure consisted of vocabulary items. The negative coefficients for this measure on the first, fourth and fifth factors are not so large as to suggest an important negative association between risk taking behavior and the abilities defined by these factors.

Conclusions

The main conclusion of this study concerns the equivalence of measurements arising from tests based on the same content but employing different formats. When content was held constant and allowance was made for differences due to errors of measurement and scale parameters, i.e., units and origins, the tests of mathematical reasoning that were employed were equivalent regardless of format, but the tests of verbal comprehension were not. In particular, the free-response tests of verbal comprehension seemed to measure something different from standard multiple-choice and Coombs tests of this ability, although the standard multiple-choice and

20

Coombs formats themselves yielded equivalent measures of verbal comprehension. This finding, if found to be generally true, has obvious methodological implications for educational and psychological researchers. The design of instruments to measure verbal comprehension must be done with the full awareness that different formats may well yield measures of different abilities. This same concern is apparently not necessary for tests of mathematical reasoning.

The foregoing, major conclusion of the study cannot go unqualified. In this study, all the pairs of tests with the same format, regardless of whether the content consisted of mathematics questions or vocabulary items, were not statistically parallel (i.e., they had different means, variances, reliability coefficients, and intercorrelations with other variables). Further evidence of the lack of parallelism was obtained from the factor analyses that were attempted, in that a parallel forms factor structure did not provide a satisfactory fit to the matrix of intercorrelations among the 12 mathematical reasoning and verbal comprehension tests. The results of the only factor analysis that gave satisfactory results indiciate that the unique factor, including error of measurement, for one form of a test was correlated with the unique factor for the "parallel" form of the test. The statistical test of equivalence provided by Lord assumes the existence of "replicate" measurements--truly parallel tests would provide replicate measurements--having errors of measurement that are uncorrelated across replications (Lord, 1971, p. 2). Nothing seems to be known about the robustness of Lord's test when this assumption is violated.

The second, and very much weaker conclusion of this study, is that evidence was obtained of the existence of a constructed-response format factor. The evidence for this factor is weak because, although the constructed response test of mathematical reasoning and verbal comprehension had, as expected, positive coefficients on this factor, all the coefficients were small in absolute magnitude and the factor did not have the expected associations with the marker variables.

The primary reason for undertaking the study, to identify format factors and gain an understanding or explanation of these factors by relating them to marker variables for following-directions ability, recall and recognition memory, and risk taking, appears to have been unjustified. It was not possible to identify format factors that were clearly marked and that accounted for a substantial amount of variance common to the tests having the same format regardless of content.

Reference Note

McDonald, R. P., & Leong, K. COSA-I program. Toronto: Ontario Institute

for Studies in Education, 1975.

References

Canadian New Achievement Test in Mathematics. Toronto: Ontario
Institute for Studies in Education, 1965.

Cook, D. L. An investigation of three aspects of free-response and choice
type tests at the college level. (Doctoral dissertation, Ohio State
University) Ann Arbor, Mich.: University Microfilms, 1955, No. 12,
886.

Coombs, C. H., Milholland, J. E., & Womer, F. B. The assessment of partial
knowledge. Educational and Psychological Measurement, 1956, 16, 13-37.

Cooperative School and College Ability Tests. Princeton, N.J.: Educational
Testing Service, 1957.

Davis, F. B., & Fifer, G. The effect on test reliability and validity of
scoring aptitude and achievement tests with weights for every choice.
Educational and Psychological Measurement, 1959, 19, 159-170.

Dressel, P. L., & Schmid, J. Some modifications of the multiple-choice
item. Educational and Psychological Measurement, 1953, 13, 574-595.

Duncanson, J. P. Learning and measured abilities. Journal of Educational
Psychology, 1966, 57(4), 220-229.

French, J. W., Ekstrom, R. B., & Price, L. A. Kit of reference tests for
cognitive factors (rev. ed.). Princeton, N.J.: Educational Testing
Service, 1963.

Hambleton, R. K., Roberts, D. M., & Traub, R. E. A comparison of the
reliability and validity of two methods for assessing partial know-
ledge on a multiple-choice test. Journal of Educational Measurement,
1970, 7, 75-82.

Heim, A. W., & Watts, K. P. An experiment on multiple-choice versus open-ended answering in a vocabulary test. British Journal of Educational Psychology, 1967, 37, 339-346.

Jöreskog, K. G. A general approach to confirmatory maximum likelihood factor analysis. Psychometrika, 1969, 34, 183-202.

Jöreskog, K. G. Statistical analysis of sets of congeneric tests. Psychometrika, 1971, 36, 109-133.

Lord, F. M. A significance test for the hypothesis that two variables measure the same trait except for errors of measurement. Psychometrika, 1957, 22, 207-220.

Lord, F. M. A study of speed factors in tests and academic grades. Psychometrika, 1956, 21, 31-50.

Lord, F. M. Testing if two measuring procedures measure the same psychological dimension. Princeton, N.J.: Educational Testing Service, Research Bulletin RB-71-36, 1971.

Lucas, C. M., & French, J. W. A factorial study of experimental tests of judgment and planning. Princeton, N.J.: Educational Testing Service Research Bulletin 53-16, 1953.

McDonald, R. P. A generalized common factor analysis based on residual covariance matrices of prescribed structure. British Journal of Mathematical and Statistical Psychology, 1969, 22, 149-163.

McNemar, Q. Attenuation and interaction. Psychometrika, 1958, 23, 259-266.

Rippey, R. Probabilistic testing. Journal of Educational Measurement, 1968, 5, 211-215.

Swineford, F. The measurement of a personality trait. Journal of Educational Psychology, 1938, 29, 295-300.

Swineford, F. Analysis of a personality trait. Journal of Educational Psychology, 1941, 32, 438-444.

Traub, R. E. A factor analysis of programmed learning and ability measures.

Canadian Journal of Behavioral Science, 1970, 2, 44-59.

Traub, R. E., & Hambleton, R. K. The effect of scoring instructions and

degree of speededness on the validity and reliability of multiple-

choice tests. Educational and Psychological Measurement, 1972, 32,

737-757.

Vernon, P. E. The determinants of reading comprehension. Educational

and Psychological Measurement, 1962, 22, 269-286.

Villegas, C. Confidence region for a linear relation. Annals of

Mathematical Statistics, 1964, 35, 780-788.

Footnotes

[1]The authors are indebted to a large number of people who helped in one way or another to make the study possible: Mary Cockell, Liz Falk, Colin Fraser, Mohindra Gill, Lorne Gundlack, Gladys Kachkowski, Kuo Leong, Joyce Townsend, Pat Tracy, Wanda Wahlstrom, Dawn Whitmore.

[2]This project was supported by a research grant from the Office of the Coordinator of Research and Development, OISE.

[3]The ability to follow directions has been called integration and defined as the "ability simultaneously to bear in mind and to combine or integrate a number of premises, or rules, in order to produce the correct response" (Lucas & French, 1953, p. 3).

[4]All ETS items were used with the full knowledge and permission of the publisher.

[5]These materials were used with permission.

[6]We are grateful to Mr. Gordon Brown, Principal, St. Clair Junior School, and Mr. Frank Gould, Principal, Westwood Junior School, their teaching staffs and students for cooperating in the study.

[7]The strategy of testing all possible pairs of instruments for equivalence can be criticized because the tests that are made are not linearly independent. In this specific instance, however, a better strategy, one that would avoid this criticism, did not suggest itself.

[8]Interested readers can obtain the statistics that were computed at each stage of each application of Lord's test of equivalence by writing to the senior author.

[9]The frequency of partial knowledge responses might reasonably be expected to increase as test difficulty increased. Differences in test difficulty does not, on average, appear to account for the present results. The mean scores on the multiple-choice versions of the mathematical reasoning and verbal comprehension tests are approximately equal to one-half the total number of items in the tests.

Table 1

Basic Information on 19 Measures: Administration Times,
Number of Items, Means, Standard Deviations, Alpha Coefficients, and Intercorrelations

Measure	Administration Time (Minutes)	No. of Items	Mean	S.D.	a	No. of Measure	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Mathematical Reasoning																							
Constructed-Response																							
– Form A	45	30	15.2	4.9	.85	1																	
– Form B	45	30	16.4	5.3	.86	2	.73																
Standard Multiple-Choice																							
– Form A	45	30	14.8	5.2	.77	3	.69	.71															
– Form B	45	30	15.5	6.0	.84	4	.58	.72	.63														
Coombs																							
– Form A	45	30	40.0	31.7	.85	5	.57	.62	.69	.66													
– Form B	45	30	45.7	33.1	.87	6	.67	.72	.69	.79	.68												
Verbal Comprehension																							
Constructed-Response																							
– Form A	30	50	15.8	8.1	.91	7	.50	.46	.44	.48	.36	.41											
– Form B	30	50	15.6	7.8	.91	8	.49	.47	.48	.45	.41	.42	.86										
Standard Multiple-Choice																							
– Form A	30	50	27.0	9.0	.89	9	.51	.47	.51	.51	.42	.46	.83	.78									
– Form B	30	50	25.0	8.9	.88	10	.48	.48	.46	.50	.43	.43	.85	.87	.86								
Coombs																							
– Form A	30	50	85.4	46.5	.90	11	.52	.49	.52	.52	.44	.48	.84	.79	.87	.87							
– Form B	30	50	74.1	45.3	.89	12	.50	.51	.50	.58	.46	.54	.82	.81	.81	.90	.84						
Following Directions, Test 1	10	11	7.4	2.2	.50	13	.38	.44	.45	.44	.40	.45	.46	.41	.41	.45	.47	.46					
Test 2	10	11	6.8	2.4	.43	14	.43	.50	.51	.55	.41	.52	.50	.48	.53	.54	.56	.53	.64				
Recall Memory, Picture-Number Test	7	21	9.8	4.5	.16	15	.16	.25	.16	.23	.25	.18	.07	.13	.05	.07	.06	.08	.27	.25			
Object-Number Test	7	15	6.3	3.7	.13	16	.13	.16	.12	.04	.23	.11	-.03	.01	-.04	-.04	-.04	-.01	.05	.03	.59		
Recognition Memory, Names	4	24	16.3	4.6	.20	17	.20	.24	.32	.28	.37	.29	.29	.32	.22	.29	.28	.29	.35	.29	.03	.02	
Figures	4	16	11.4	3.5	.19	18	.19	.13	.25	.15	.16	.20	.23	.19	.14	.21	.22	.22	.23	.13	-.04	-.08	.36
Risk Taking	45	78	15.7	19.4	.04	19	.04	.12	.04	.06	.07	-.01	.17	.15	.17	.18	.17	.18	.08	.07	.15	.12	.12

Table 2

Means and Standard Deviations of the Distributions of the
Frequencies with which Students Employed each Type of Response in the Coombs
Format (Standard deviations appear in brackets)

Response	Associated Score	Mathematical Reasoning		Verbal Comprehension	
		Form A	Form B	Form A	Form B
a) 4 wrong options	4	13.5 (6.4)	14.5 (6.8)	23.6 (10.8)	21.0 (10.3)
b) 3 wrong options	3	.5 (1.0)	.6 (1.2)	3.7 (4.6)	4.2 (4.7)
c) 2 wrong options	2	.4 (1.0)	.6 (1.6)	1.5 (2.5)	1.9 (3.1)
d) 1 wrong option	1	.5 (1.9)	.3 (1.0)	1.1 (2.7)	1.0 (3.0)
e) 0 wrong options	0	1.0 (2.2)	.7 (1.4)	1.0 (2.6)	1.0 (2.8)
f) 3 wrong options plus correct answer	-1	12.9 (6.0)	12.1 (6.6)	15.7 (8.7)	16.6 (10.0)
g) 2 wrong options plus correct answer	-2	.7 (1.6)	.7 (1.6)	2.2 (3.0)	3.1 (4.2)
h) 1 wrong option plus correct answer	-3	.3 (.8)	.4 (1.1)	.7 (1.5)	.9 (1.6)
i) 0 wrong options plus correct answer	-4	.1 (.4)	.1 (.4)	.2 (.6)	.2 (.6)
j) All options marked	-5	.3 (.9)	.1 (.5)	.2 (.6)	.2 (.5)
Percentage of partial knowledge responses[a]		8.3	9.0	18.8	22.5

_____

[a]Computed from the formula: $\dfrac{\text{(Sum of means for responses b, c, d, g, h, i)}}{\text{Sum of means for all responses}} \times 100$.

The sum of means for all responses differs from 30 or 50 because of rounding error.

Table 3

Estimated Coefficients of a Five-Factor Structure Fitted to the Matrix of Intercorrelations Among 12 Mathematical Reasoning and Verbal Comprehension Tests

Structural Equation: $R^* = H\,S\,H' + U$

where R^* is the structural approximation to R, the intercorrelation matrix formed from the first 12 rows and columns of the intercorrelation matrix reported in Table 2.

H =

VARIABLE	I	II	III	IV	V
1. MR (A) – CR	.81		.22		
2. MR (B) – CR	.78		.11		
3. MR (A) – SMC	.81			.21	
4. MR (B) – SMC	.88			-.02	
5. MR (A) – CMC	.71				.01
6. MR (B) – CMC	.76				.32
7. VC (A) – CR		.75	.25		
8. VC (B) – CR		.75	.13		
9. VC (A) – SMC		.82		.06	
10. VC (B) – SMC		.80		-.12	
11. VC (A) – CMC		.85			.04
12. VC (B) – CMC		.88			.19

(FACTORS)

U =

```
 .30   .07   .05  -.13   .00   .06
 .07   .38   .09   .03   .07   .13
 .05   .09   .31  -.07   .12   .08
-.13   .03  -.07   .23   .04   .12
 .00   .07   .12   .04   .50   .13
 .06   .13   .08   .12   .13   .32
                                      .38   .27   .21   .25   .20   .16
                                      .27   .43   .17   .27   .15   .16
                                      .21   .17   .32   .21   .17   .08
                                      .25   .27   .21   .34   .18   .20
                                      .20   .15   .17   .27   .18   .08
                                      .16   .16   .08   .20   .08   .19
```

S is the 5x5 identity matrix except for the correlation between factors I and II which is estimated to be 0.73

Note: The blank cells of H and U are zeros.