EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
1993, 53

MORE EVIDENCE IN FAVOR OF THREE-OPTION MULTIPLE-CHOICE TESTS

R. ERIC LANDRUM
Boise State University

JEFFREY R. CASHIN AND KRISTINA S. THEIS
University of Wisconsin-Platteville

Students from two consecutive semesters were given multiple-choice tests over five units of an undergraduate course in psychology. During the first semester, students were given five 50-question 4-option multiple-choice tests, and during the second semester students were given five 50-question 3-option multiple-choice tests. One-hundred and forty-four (57.6%) of the questions were identical between semesters except for second semester test items having only 3 options. Results indicate that students performed significantly better on 3-option items than on 4-option items (corrected for chance guessing), and that this improvement may be due to improved validity of the test items.

THE study of multiple-choice testing has enjoyed a long history. Almost since the time the multiple-choice test was developed by Otis in the late 1910s (see Boring, 1950), measurement specialists have been interested in the tests' ability to provide both reliable and valid results. Lord (1977) reported that in the 1920s, systematic studies were being conducted by Toops (1921) and Ruch and colleagues (Ruch and Charles, 1928; Ruch, DeGraff, and Gordon,

Various portions of this report were previously presented at the Tri-State Undergraduate Psychology Conference held in Platteville, WI November 1991 and at the Midwestern Psychological Association meeting held in Chicago May 1992.

Address any correspondence (including requests for reprints) to R. Eric Landrum, Department of Psychology, Boise State University, 1910 University Drive, Boise, ID 83725.

1926; Ruch and Stoddard, 1925; 1927) on the effectiveness of multiple-choice testing. Of interest then and now is the issue of how to optimize the value of multiple-choice tests. Given the desire to assess knowledge in a fair and equitable manner, can the multiple-choice test satisfy these goals?

Much of the modern work in the field followed a seminal work by Tversky (1964). Tversky suggested by means of a mathematical proof that the power of a test, its ability to discriminate between test takers, and the amount of information a test can provide are factors that are all optimized in a 3-option test compared to other types of tests when the total number of alternatives is fixed. An example of satisfying Tversky's criterion would be in the comparison of a 30-question 4-option test with a 40-question 3-option test; both tests contain 120 options (i.e., fixed).

Other studies conducted on a theoretical basis have concluded that the 3-option test is superior to the 4- and 5-option tests. Lord (1977) came to a similar conclusion about 3-option tests, but added that the change is to the advantage of high-level students and to the disadvantage of low-level students (however, see Trevisan, Sax, and Michael [1991] for differing results). Mattson (1965), Ebel (1969), and Grier (1975) also hypothesized that the change from a 4-option item to a 3-option item increased the reliability of the test.

While the theoretical evidence clearly demonstrates the advantages of the 3-option tests, there have only been a few empirical studies reported in the literature implementing the 3-option test. Generally, the studies have shown a varied amount of support for the 3-option multiple-choice item, citing Tversky's (1964) advantages in addition to demonstrations of improved reliability (Costin, 1970; 1972; Hogben, 1973; Owen and Froman, 1987; Stratton and Catts, 1980; Williams and Ebel, 1957). Other benefits commonly emerge with the 3-option test, such as the increased ease of item generation by the instructor, students answering questions with less distractions, and faster completion of the items by students, allowing instructors to test more concepts (Costin, 1970; Grier, 1975; Owen and Froman, 1987).

While the accumulation of evidence suggests that 3-option items are often superior to 4-option items, conventional wisdom still suggests that each item have four choices. This was evidenced by Owen and Froman's (1987) analysis of measurement textbooks with an overwhelming bias for the 4-option item, without providing empirical support for its' use. Additionally, most testbanks now available with popular introductory textbooks are composed of predominantly 4-option multiple-choice questions. Owen and Fro-

man (1987) concluded this most recent examination of 3-option items by suggesting that the items be studied in the context of "typical classroom tests built by teachers" (p. 520). The present study examined the transition from 4-option items to 3-option items in an introductory psychology course.

In making this transition from 4-option to 3-option items, the major concerns of the present study were: (a) how will the change affect student grades, and (b) will the 3-option test serve as a better test of students' knowledge? If a test can be designed to be a better test of student knowledge, one would expect some grades to increase and others to decrease. However, changes in student performance can result from other factors such as asking easier questions (e.g., more cues or poor item construction) or harder questions (e.g., providing no contextual cues). The desired outcome of the present study would result in a 3-option test format without dramatic changes in difficulty, perhaps giving the student greater opportunity to demonstrate their knowledge of the subject matter. Other benefits certainly motivate the transition from 4-option to 3-option items, such as the savings in time by students taking the test, instructors constructing the test, and the opportunity for instructors to ask more questions.

Method

Subjects

Subjects were undergraduate students enrolled in a General Psychology course at the University of Wisconsin-Platteville. During Semester 1 (Fall 1990), 46 students completed the course, and during Semester 2 (Spring 1991) 67 different students completed the course. This course fulfills a general education requirement, with a cross section of majors enrolled in both semesters.

Materials

Test items were a mixture of instructor-generated multiple-choice questions and publisher-supplied questions. As the course was divided into units, items were selected from all units of the course (see Table 1 for the unit titles and the number of test items). During Semester 1, 4-option tests were given, and during Semester 2, 3-option tests were given. Four-option items were modified into 3-option items by the course instructor. This was a subjective

TABLE 1
Unit Organization of General Psychology Course with Number of Shared Questions

Unit No.	Unit Name	No. of Questions in Common
1	Introduction/Research Methods	25
2	Learning, Memory, Cognition, Language	22
3	Biological Bases of Behavior	16
4	Sensation and Perception	38
5	Social Psychology	43
	Overall	144

Note. For both semesters, 50-question tests were given. The above numbers represent how many 4-option items were reused the following semester as 3-option items.

judgment completed by eliminating the least plausible option from each 4-option item.

Procedure

Students in both General Psychology classes completed the course organized around five units of study (see Table 1). Semester 1 students received a 50-question 4-option multiple-choice test at the conclusion of each unit. Semester 2 students also received a 50-question multiple-choice test at the conclusion of each unit, but the test items contained only three options. The instructor and lecture content remained constant (as much as possible) over the course of the academic year (1990–1991). However, identical tests were not given over the course of both semesters. While that method would have been desirable from a research standpoint, from the real classroom standpoint giving the exact same test (except for number of options) might disadvantage those first semester students.

Only a limited subset (144 of 250, or 57.6%) of the items from the 4-option tests were reused in 3-option tests in the second semester, and the number of common items in each unit can be found in Table 1. In other words, 144 items were exactly repeated across Semesters 1 and 2, with the Semester 2 version items only having three options. Subsequent analyses examining student performance on the test items are based solely on the question set in common across both semesters, thus equalizing the conceptual difficulty of test item sets (except for any difficulty differences due to number of options).

Design

Given the limitations of the actual classroom situation, emphasis was placed on how students responded to particular test items. In

TABLE 2
Item Difficulty (D) Performance on Multiple Choice Tests (Percentage Correct)

Unit No.	Semester 1 4-option	Semester 2 3-option	*t* test	*p*
1	80.1	88.0	$t(24)=-3.94$	<.001
2	76.5	83.9	$t(21)=-3.20$	<.005
3	82.9	91.4	$t(15)=-3.45$	<.005
4	83.1	84.9	$t(37)=-1.21$	n.s.
5	84.8	87.7	$t(42)=-2.00$	.051
Overall	82.0	86.8	$t(143)=-5.70$	<.0001

Note. The higher the score, the better the test performance.

other words, the items served as the unit of analysis for this study. Unit-by-unit as well as overall analyses from both courses were conducted comparing 4-option vs. 3-option test performance using paired *t* tests (Owen and Froman, 1987).

Results

The two major concerns of the present study were (a) student performance on the test items, and (b) how well the tests were able to capture student knowledge of the subject matter. The comparisons between 4- and 3-option performance were made on an item-by-item basis.

Student Performance

The unit-by-unit breakdown and overall performance measures are presented in Table 2. Table 2 presents the mean scores on the common items as well as the paired *t* test results. In examining student performance on the common items between Semester 1 (4-option) and Semester 2 (3-option), students scored significantly higher on the Unit 1, 2, 3, and 5 tests, and the overall comparison of items found that students scored significantly higher grades on the 3-option items.

Test-Item Performance

Another concern related to student performance is that of the performance of the test items. In other words, were the items adequate in assessing student knowledge? Given the general increase in student performance with 3-option items, does the change mean that the test is easier (less difficult), or did the test just do a

TABLE 3
Modified Item Difficulty (D') Mean Scores (Corrected for Chance Guessing)

Unit No.	Semester 1 4-option	Semester 2 3-option	*t* test	*p*
1	55.2	55.0	$t(24)=+0.09$	n.s.
2	51.5	50.9	$t(21)=+0.23$	n.s.
3	57.9	58.4	$t(15)=-0.21$	n.s.
4	58.1	51.9	$t(37)=+3.92$	<.0005
5	59.8	54.7	$t(42)=+3.47$	<.005
Overall	57.0	53.8	$t(143)=+3.77$	<.0005

Notes. Modified difficulty was calculated for each item by subtracting the probability of chance guessing from the percentage correctly responding. The lower the score, the more difficult a test.

better job at capturing student knowledge? In order to pursue this notion, a modification of the standard difficulty score was made. Difficulty, referred to here as *D*, is traditionally calculated as the percentage of students responding correctly to an item.

The modified version of difficulty, *D'*, was calculated by subtracting the chance probability of getting an item correct from the actual percentage of correct responses. Thus for 4-option items, $D' = D\text{-}25\%$, and for 3-option items, $D' = D\text{-}33\%$. Because the probability of chance guessing improves from 25% to 33% with the change from 4-option to 3-option items, this correction to difficulty (*D'*) helps to equalize the scores and provide a better comparative picture of student performance. The *D'* scores and accompanying statistical analyses are presented in Table 3. While Units 1, 2, and 3 did not differ significantly after correction, Units 4 and 5 and the overall analysis indicated significant differences in difficulty. In this latter grouping, the 3-option items were more difficult after correcting for guessing than the 4-option items.

Discussion

What is the effect of changing from a 4-option multiple-choice test to a 3-option multiple-choice test in the classroom? This study approached that issue from two viewpoints: (a) student performance and (b) test-item performance.

Students' scores on the common items rose significantly on all tests except one (Unit 4). The general pattern of results represented in the overall category suggest that students get higher grades when taking 3-option multiple-choice tests compared to those students taking 4-option tests.

However, instructors are also concerned with how well the test

measures student knowledge. For example, is the 3-option test a better test or just an example of grade inflation? An indirect answer to this question comes from the examination of corrected difficulty scores. While three units of the course experienced no significant change in difficulty, two units and the overall measure indicated that the 3-option test was actually more difficult, evidenced by lower difficulty measures. Student performance increased while at the same time test-item difficulty also increased (or at least remained constant).

These seemingly paradoxical results do appear consistent with previous findings and also with the conclusion that the 3-option test is a better measure of student knowledge. Students are less distracted in the 3-option situation and can complete the items more quickly than with four options (Owen and Froman, 1987; Stratton and Catts, 1980). While the proportion of guessing increases with the 3-option items, correcting for guessing makes the tests look as difficult or slightly more difficult (from a statistical standpoint) as before. However, when students have fewer distractions, the beneficial effects predicted by Tversky (1964), Grier (1975), and Lord (1977) do seem to occur.

The present study concludes as other before it (Costin, 1970; 1972; Hogben, 1973; Owen and Froman, 1987; Stratton and Catts, 1980; Williams and Ebel, 1957) that the use of 3-option multiple-choice item is the preferred method of multiple-choice testing. Students seem to prefer the 3-option test because they score higher on it and perceive it to be less confusing and less tricky (Owen and Froman, 1987). The present study provides some indirect evidence that students perform better on the 3-option test even though that test remains as difficult or more difficult. The 3-option test appears to be a better, perhaps more valid test of student knowledge, and this improved testing format resulted in enhanced student performance in this study.

REFERENCES

Boring, E. G. (1950). *A history of experimental psychology* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Costin, F. (1970). The optimal number of alternatives in multiple-choice achievement 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. L. (1969). Expected reliability as a function of choices per item. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 29, 565–570.

Grier, J. B. (1975). The number of alternatives for optimum test reliability. *Journal of Educational Measurement*, 12, 109–112.

Hogben, D. (1973). The reliability, discrimination and difficulty of word-knowledge tests employing multiple choice items containing three, four, or five alternatives. *The Australian Journal of Education*, 17, 63–68.

Lord, F. M. (1977). Optimal number of choices per item—a comparison of four approaches. *Journal of Educational Measurement*, 14, 33–38.

Mattson, D. (1965). The effects of guessing on the standard error of measurement and the reliability of test scores. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 25, 727–730.

Owen, S. V. and Froman, R. D. (1987). What's wrong with three-option multiple choice items? EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 47, 513–522.

Ruch, G. M. and Charles, J. W. (1928). A comparison of five types of objective tests in elementary psychology. *Journal of Applied Psychology*, 12, 398–404.

Ruch, G. M., DeGraff, M. H., and Gordon, W. E. (1926). *Objective examination methods in the social studies.* New York: Scott Foresman.

Ruch, G. M. and Stoddard, G. D. (1925). Comparative reliabilities of five types of objective examinations. *Journal of Educational Psychology*, 16, 89–103.

Ruch, G. M. and Stoddard, G. D. (1927). *Tests and measurements in high school instruction.* Chicago: World Book.

Stratton, R. G. and Catts, R. M. (1980). A comparison of two, three and four-choice item tests given a fixed total number of choices. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 40, 357–365.

Toops, H. (1921). Trade tests in education. In *Teachers College Contributions to Education* (No. 115). New York: Columbia University.

Trevisan, M. S., Sax, G., and Michael, W. B. (1991). The effects of the number of options per item and student ability on test validity and reliability. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 51, 829–837.

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

Williams, B. J. and Ebel, R. L. (1957). The effect of varying the number of alternatives per item on multiple-choice vocabulary test items. In *The Fourteenth Yearbook of the National Council on Measurements Used in Education* (pp. 63–65). East Lansing, MI: Michigan State University.