JOURNAL OF EDUCATIONAL MEASUREMENT
VOLUME 12, NO. 4 WINTER 1975

A COMPARISON OF SEVERAL METHODS OF ASSESSING PARTIAL KNOWLEDGE IN MULTIPLE-CHOICE TESTS: II. TESTING PROCEDURES*

A. RALPH HAKSTIAN
University of British Columbia
and
WANLOP KANSUP
University of Alberta

Several testing procedures for assessing partial knowledge in multiple-choice tests have been proposed over the years. As noted by Wang and Stanley (1970), they have been based on a concept of "response" to an item that is broader than simply indicating the correct alternative. Among many measurement specialists, considerable optimism, regarding improved reliability and validity, remains for these procedures (see, e.g., Collet, 1971; Hambleton, Roberts, & Traub, 1970). Two variables are operative in generating different sets of test scores: (1) the method by which examinees respond, and (2) the method by which the obtained responses are scored. In an investigation of different item-response instructions, the final product evaluated results from an interaction of response methods and scoring methods. The authors have attempted to separate these two factors, and, in another paper (Kansup & Hakstian, 1975), have dealt with scoring procedures. With the results of this earlier study established, the authors' concern in the present paper is with different item-response methods.

The two best-known testing procedures aimed at assessing partial knowledge and reducing the effects of guessing in multiple-choice tests appear to be *elimination testing* (Coombs, Milholland, & Womer, 1956), and *confidence,* or *probabilistic testing* (de Finetti, 1965; Dressel & Schmid, 1953; Hambleton et al., 1970; Hopkins, Hakstian, & Hopkins, 1973; Koehler, 1971; Michael, 1968; Rippey, 1968, 1970; Shuford, Albert, & Massengill, 1966).

Dressel and Schmid (1953), using achievement tests, examined an early variant of elimination testing, referred to as the "free-choice" method. Examinees marked as many options as needed to ensure that the correct answer was not omitted. Each incorrect alternative marked was assigned a score of -¼, with the correct option scored +1. This scoring resulted in an insignificant gain in internal consistency, compared to conventional scoring; validity was not examined. Coombs et al. (1956), using several aptitude tests, had examinees cross out those of the four options they believed to be incorrect. They assigned scores of +1 for each incorrect alternative crossed out, and -3 for a correct option crossed out. A small, statistically insignificant gain in internal consistency was found for these scores, over those of a conventionally-tested group; again validity was not examined. Collet (1971), using IQ tests, found no significant gain in reliability of scores for a group responding by the method of Coombs et al. (1956), compared to scores of a conventionally-tested group, but he did point to a significant gain in

*This study was supported by aid from the Canadian International Development Agency. The authors are grateful to Drs. S. M. Hunka, T. O. Maguire, and V. R. Nyberg, and to Murray Richmond, Raymond Baril, and Ernest Skakun for assistance in various stages of the study.

criterion-related validity. However, a comparison of the validities for the two groups, using the usual z-test for Fisher's Z-transformed independent correlations, reveals that the difference failed to reach even the .10 level of significance. In past research with the elimination method, then, no substantial gain in reliability has resulted, and the evidence regarding validity is, at present, provisional.

In confidence testing, the examinee indicates his degree of confidence, or subjective probability, that the option he chooses (or in some cases, each option) is correct. Dressel and Schmid (1953) employed a "degree of certainty" response procedure with achievement tests, in which examinees expressed their confidence, on a scale of 1 to 4, in each response. An insignificant gain in internal consistency was obtained for these scores, compared to conventional scores. Hopkins et al. (1973) employed a similar response method, using a 3-point scale with an achievement test, and found a slight, but nonsignificant, increase in internal consistency for the confidence-weighted responses, and a slight, but nonsignificant, *decrease* in validity. Using a 10-point confidence distribution method, Michael (1968) obtained a gain in internal consistency, for the confidence scores, but this was not assessed for statistical significance. Using more sophisticated confidence-responding procedures and scoring functions, Rippey (1968), Hambleton et al. (1970), and Koehler (1971) found no gain in internal consistency for confidence tests compared to conventional tests. In the latter two studies (both employing achievement tests) validity was examined, and no significant increases in validity were obtained. The disappointing distillation of the past research with confidence testing, then, contains no evidence of a reliable increase in either reliability or validity, when compared to conventional testing.

In the present study, an attempt was made to evaluate comprehensively the elimination and confidence methods of testing, using conventional testing as the baseline for comparisons. A large sample was randomly divided into three groups, each receiving one test-taking "treatment." Some variety in test content was achieved. Reliability was treated more fully than in past studies, by an examination of both internal consistency and stability. A broad examination of criterion-related validity was provided by use of a variety of school achievement criteria and of similar ability measures. The *scoring* of the confidence tests was accomplished using the procedure found in the authors' earlier study (Kansup & Hakstian, 1975) to be both simplest, and generally, at least as reliable and valid as the other methods compared.

METHOD

Subjects

A total of 1028 grade nine students, on whom complete data were ultimately available, were randomly assigned to three test-taking groups: (1) the *C* group: 346 students (with the sexes about equally represented) given tests with conventional instructions, (2) the *CW* group: 348 students given the same tests but with confidence-weighting instructions, and (3) the *EL* group: 334 students given the tests with elimination-testing instructions. Eight junior high schools in the greater Edmonton, Alberta area participated in the study, and random assignment was made *within* each of 44 classes. Thus, roughly one third of each class was assigned to each group. The schools chosen were relatively homogeneous in social class, and would be characterized as lower-middle to middle class. The students did not volunteer; instead whole class periods were set aside for the study.

Instruments

Two Verbal Ability tests containing vocabulary items, designated VA and VB, and two Mathematical Reasoning tests (containing numerical problems), designated RA and RB, were compiled from the *Kit of Reference Tests for Cognitive Factors* (French, Ekstrom, & Price, 1963). The VA test contained 25 five-choice items; the VB test contained 30 four-choice items. The RA and RB tests each contained 15 five-choice items. The means and standard deviations, respectively, in the *C* group were VA: 12.0, 3.5; VB: 16.0, 5.1; RA: 6.7, 2.7; RB: 5.6; 2.7.

A preliminary study involving 28 grade nine students not included in the full analysis was carried out to establish time limits for all instruments that would permit virtually everyone in the larger study to finish all tests. The VA and RA tests were administered to these preliminary examinees under the *most time-consuming* testing conditions of the three investigated (*confidence* testing). The times when all 28 preliminary examinees had completed the VA and RA tests were noted, and these times (12 minutes for the VA test; 15 minutes for the RA) were used as time limits in the larger study. Since, in the full study, the VB and RB tests were administered to all students under *conventional* conditions, they were similarly administered to the 28 preliminary examinees, and the times when all 28 students had finished these tests were again noted. These times (VB, 8 minutes; RB, 10 minutes) were subsequently used as time limits in the larger study. Speededness, therefore, was not a factor in this study.

To assess validity, semester-end grades were obtained in Language Arts, Mathematics, and Science courses, as were overall grade averages. These four measures were standardized *within* each school group and then pooled across schools.

Testing Procedures and Scoring

Students in the three groups took the tests together in their regular classrooms. First, a general orientation was provided, in which the students were briefed on test booklets, answer sheets, etc., and were informed that some of them would be taking different kinds of tests than others. Students then *read* to themselves detailed instructions specific to their group, prior to responding to the items. Several examples were provided for students in each group. After the students had read their particular instructions, a question period was provided to ensure that each student fully understood the instructions specific to his or her group. There were very few questions in any of the classrooms: the written instructions were generally adequate in conveying the different item-response requirements. In the few instances in which a student indicated a failure to understand the instructions; the examiner went to the student's desk and quietly cleared up the difficulty, so as not to confuse students in the other test-taking groups. During the testing, students were generally unaware of the method of responding in groups other than their own.

Inspection of completed answer sheets from all three groups revealed that the students had, without exception, correctly employed the item response technique for their group. Although students in all three groups were instructed to answer *every item* of each test, over the six test administrations—two each of VA and RA and one each of VB and RB—about 10 percent of the students in each group omitted one or more items on at least one test. These students' results were excluded from the subsequent analyses, so that this study included only those students who had responded to every item on every test. The final *n*'s were those given earlier: for the *C* group, 346; for the *CW*, 348; and for the *EL*, 334. Thus, the problem of possible differential effects of item omissions under the testing procedures examined was eliminated as a potentially confounding factor.

Specific details of the testing instructions and scoring procedures follow.

1. *Conventionally-Tested (C) Group.* On all tests, students in this group were given conventional instructions: Fill in the slot on the answer sheet corresponding to the correct alternative. These students were told that they should answer every item, guessing when unsure of the correct answer. Lord (1975) referred to such instructions as *number-right scoring directions.* Items were scored 0–1, and since only subjects for whom no omissions were found were included in the analyses, the obtained scores for the *C* students were identical to what would have been obtained with any formula scoring technique, including that discussed by Lord (1975).

2. *Confidence-Tested (CW) Group.* On the VA and RA tests, students in this group were instructed to distribute—on a special answer sheet—10 points of confidence among the alternatives for each item, using whole numbers only. Students were informed that to maximize their scores they would have to provide an honest indication of their confidence in each alternative. Although the confidence was thus apportioned in coarser quantities than in the more nearly continuous procedures of, for example, Rippey (1968, 1970) and Hambleton et al. (1970), the present procedure (used also by Michael, 1968) appeared to the authors to accomplish the same purpose, and be more conveniently used. The *CW* students were also instructed to answer every item.

The scoring of the responses was done five different ways, and a comparison of these different scoring procedures appears in the authors' earlier paper (Kansup & Hakstian, 1975). It was found in the earlier study that, of the five scoring methods, the simplest—taking as an item score the number of points of confidence assigned to the keyed alternative—was generally at least as reliable and valid as the more complicated, nonlinear functions compared. In particular, a *reproducing scoring system* (Shuford et al., 1966)—that is, a scoring procedure for confidence tests that permits examinees to maximize their expected item scores only by honestly assigning their confidence—was uniformly less reliable and valid (though not significantly so) than the simpler function examined. In the present study then, all results for the *CW* group are based on the simple scoring procedure noted above. It will be seen from the results that the main conclusions drawn from the present study regarding confidence testing apply a fortiori to confidence testing accompanied by the (logarithmic) reproducing scoring procedure. A more detailed discussion of the relationship between confidence testing instructions and scoring systems appears in the authors' earlier paper (Kansup & Hakstian, 1975).

3. *Elimination-Tested (EL) Group.* On the VA and RA tests, students in this group were instructed to respond on a special answer sheet by crossing out incorrect alternatives, taking care not to cross out the correct one. Students in this group were also instructed to answer every item on each test. Each incorrect option crossed out was scored + 1; a correct option, if crossed out, was scored -4. This instruction and scoring procedure was a 5-alternative version of that used by Coombs et al. (1956).

At the first of two testing sessions, in October-November, 1972, the VA and RA tests were administered, with the resulting scores denoted VA1 and RA1 in what follows. Four weeks later, all students again took the same tests with identical instructions; these scores are denoted VA2 and RA2. Also, at this second session, the VB and RB tests were given. For these tests administration and scoring was conventional for all students. Approximately 5–6 weeks after the second testing session, the various semester-end achievement measures were obtained.

Analysis Procedures

1. *Reliability.* For each group, the authors estimated (a) internal consistency from the alpha coefficients of the VA1, VA2, RA1, and RA2 scores; and (b) stability from the test-retest correlations of the VA (1 vs. 2) and RA (1 vs. 2) tests.

2. *Validity.* For each group, the authors estimated criterion-related validities of the VA1 and RA1 scores, using as criteria (a) the four school achievement measures and (b) the similar measures of Verbal and Mathematical Reasoning abilities, given by the VB and RB tests, respectively. These tests were not parallel forms of the VA and RA tests, and, in the *CW* and *EL* groups, were administered and scored differently (conventionally).

3. *Comparative Analyses.* All comparisons of reliability and validity involved *independent* groups. Pairs of independent alpha coefficients were statistically compared using Feldt's (1969) procedure. Pairs of test-retest and validity coefficients were statistically compared using the well-known two-sample z-test for Fisher's Z-transformed independent correlations.

RESULTS

Reliability

Results of all reliability assessments appear in Table 1. It is seen from Table 1 that no increase in reliability (neither internal consistency nor stability) resulted for the scores obtained by the elimination method, compared to those obtained by the conventional method. Any differences were slight and did not consistently favor one group. On the other hand, there is some evidence that confidence testing results in higher internal consistency and stability than does either conventional testing or elimination testing.

TABLE 1

Summary of Reliability Data Obtained

	Alpha Coefficients				Test-Retest Coefficients	
	Test Scores				Test	
Test-Taking Group	VA1	VA2	RA1	RA2	VA(1 vs 2)	RA(1 vs 2)
Conventional (C)	.627	.673	.596	.639	.680	.607
Confidence (CW)	.743	.722	.691	.695	.764	.669
Elimination (EL)	.647	.689	.565	.640	.671	.599
Significance of Pairwise Comparisons (p values)[1]						
C vs CW	<.01	>.10	<.05	>.10	<.05	>.10
C vs EL	>.50	>.50	>.25	>.50	>.50	>.50
CW vs EL	<.01	>.25	<.01	>.10	<.05	>.10

[1] All significance tests were two tailed. Those on pairs of independent alpha coefficients were performed using Feldt's (1969) procedure. Those on pairs of independent test-retest coefficients were performed using the z-test for Fisher's Z-transformed correlations.

TABLE 2

Summary of Validity Data Obtained

1. VA1 SCORES

	Criterion-Related Validities				
	School Achievement Criterion				Similar Ability Test[1]
Test-Taking Group	L. Arts	Math	Science	Avg.	(Verbal Ability)
Conventional (C)	.486	.445	.472	.517	.589
Confidence (CW)	.336	.333	.400	.410	.646
Elimination (EL)	.395	.350	.432	.435	.672
Significance of Pairwise Comparisons (p values)[2]					
C vs CW	<.05	<.10	>.10	<.10	>.10
C vs EL	>.10	>.10	>.50	>.10	<.10
CW vs EL	>.25	>.50	>.50	>.50	>.50

2. RA1 SCORES

	Criterion-Related Validities				
	School Achievement Criterion				Similar Ability Test[1]
Test-Taking Group	L. Arts	Math	Science	Avg.	(Math. Reasoning)
Conventional (C)	.271	.470	.395	.422	.496
Confidence (CW)	.304	.499	.426	.466	.488
Elimination (EL)	.355	.556	.405	.462	.383
Significance of Pairwise Comparisons (p values)[2]					
C vs CW	>.50	>.50	>.50	>.25	>.50
C vs EL	>.10	>.10	>.50	>.50	<.10
CW vs EL	>.25	>.25	>.50	>.50	<.10

[1] The similar ability tests were conventionally administered and scored.

[2] All significance tests were two tailed and were performed using the z-test for Fisher's Z-transformed correlations.

Validity

Results of all validity assessments appear in Table 2. No clear pattern is apparent in the validities for the *C* and *EL* groups. For the VA1 test, the conventional scores were generally more valid (though not significantly so) for the school achievement criteria, whereas for the similar Verbal Ability measure, the elimination scores were somewhat more valid. For the RA1 scores, a slight (but nonsignificant) increase in validity was obtained for the elimination scores on two of the school achievement criteria, but a *decrease* in validity for the elimination scores approaching statistical significance was found for the similar Mathematical Reasoning test. In comparing the conventional and confidence scores we see that for the VA1 scores, the conventional were clearly more valid for the achievement criteria than were the confidence scores, whereas the confidence scores were slightly (but not significantly) more valid for the similar Verbal Ability measure. For the RA1 scores, no difference in validities approached statistical significance. No significant differences in any of the obtained validities were found between the elimination and confidence scores. Clearly, neither of the experimental testing procedures—elimination testing or confidence testing—resulted in a consistent increase in validity compared to conventional testing; in several cases, a *decrease* was noted.

One seeming anomaly in Table 2 is that the validities for the similar ability measure increased somewhat for the confidence and elimination VA1 scores (although not significantly so at the .05 level), whereas the same phenomenon was not found for the RA1 scores. On might have expected the pattern found for the RA1 scores to have occurred also with the VA1 scores; there is more common *method* variance between conventional VA1 scores and VB scores (both administered and scored conventionally) than between confidence or elimination VA1 scores and VB scores. If, on the other hand, confidence and elimination testing in fact do get at partial knowledge of the ability trait measured and, hence, result in more construct valid measures of these traits, we might expect an increase in common *trait* variance between confidence and elimination VA1 scores and VB scores. This hypothesized increase in construct validity for the confidence and elimination scores should have resulted in higher criterion-related validities with the various school achievement criteria (which did not occur), since common method variance between the VA1 scores and school achievement criteria could be expected to be relatively uniform, and low, for the three testing conditions examined. The authors have no explanation for this anomaly, and subsequent research might well be directed at the effects on construct validity of experimental test-taking methods (see Koehler, 1971, for an examination of this issue with respect to confidence testing).

DISCUSSION

After some years of optimism, is would seem that the time has arrived for a critical examination of the efficacy of assessing partial knowledge in multiple-choice tests and the testing methods by which this may be done. The two best-known methods, the elimination and confidence testing procedures examined in the present study, require special training for examinees and considerably more testing time than does conventional testing. A necessary condition for the adoption of these procedures should be that reliability and validity be significantly increased over those of a conventionally-administered test requiring the same time. In the present study, neither reliability nor validity was consistently increased by the experimental methods, compared to those of conventional tests requiring the same time. In fact, validity was in some cases *decreased* for the confidence-weighted scores (notably with the VA1 scores), even without an administration time correction for the conventional test.

As noted earlier, there is no prior evidence of increased reliability for scores obtained by the elimination method (Collet, 1971; Coombs et al., 1956; Dressel & Schmid, 1953), and only tenuous evidence of increased validity (Collet, 1971). These earlier results (based on both aptitude and achievement tests), coupled with the present results (which were based on two quite dissimilar ability measures and a variety of criteria), suggest that elimination testing has little, if anything, to recommend it over conventional testing.

With regard to confidence testing, the findings are not quite so clear. Prior studies—using both aptitude and achievement tests—(Dressel & Schmid, 1953; Hambleton et al., 1970; Hopkins et al., 1973; Koehler, 1971; Michael, 1968; Rippey, 1968) have shown no significant gain in reliability for confidence tests, although two (Hopkins et al., 1973; Michael, 1968) demonstrated slight gains that were either nonsignificant or untested. In the present study, there was some evidence of an increase in both internal consistency and stability for the confidence-weighted scores compared to conventional testing. If, however, the conventionally-administered tests in the present study were lengthened by, say, 25 percent–50 percent (approximately the increase in testing time for confidence tests), the new estimates of reliability for the *C* group (from application of the Spearman-Brown formula to the coefficients in Table 1) would be equal to, or in some cases greater than, the reliabilities in the *CW* group. Thus the gains in reliability obtained in the present study are not sufficiently large to effectively favor confidence testing. In terms of validity, no prior study of confidence tests has demonstrated a significant gain over conventional testing, and the present results are consistent. In fact, significant and nearly-significant *decreases* in validity were found for confidence-weighted scores on the VA1 test using various school achievement criteria.

For those who would require confidence testing to be accompanied by a reproducing scoring system (see, e.g., Williams & Millman, 1970), the preceding conclusions apply even more strongly. When the confidence responses in the present study were scored by the logarithmic reproducing scoring system, reliabilities and validities were uniformly lower than those reported in this paper (Kansup & Hakstian, 1975).

It appears from the validity data, that by measuring subjects' confidence in their responses in addition to their grasp of item content, we measure an additional trait largely unrelated to several criteria of interest. As noted earlier by Hopkins et al. (1973), individual differences in perceived confidence may well be somewhat reliably measured; scores containing a mixture of this trait and knowledge or ability appear to be slightly more reliable than those reflecting knowledge or ability alone. However, one can use the time spent in assessing confidence to conventionally administer more knowledge or ability items. The net result will be equal reliability per unit of testing time, and, more importantly, superior validity for most criteria. On the basis of past and present empirical results—and notwithstanding elegant mathematical developments of probabilistic testing (de Finetti, 1965; Shuford et al., 1966)—the authors must conclude that, in terms of current methods of implementing it and common scholastic criteria, confidence testing, like elimination testing, appears to have little to recommend it over conventional testing.

REFERENCES

COLLET, L. S. Elimination scoring: An empirical evaluation. *Journal of Educational Measurement*, 1971, **8**, 209–214.

COOMBS, C. H., MILHOLLAND, J. E., & WOMER, F. B. The assessment of partial knowledge. *Educational and Psychological Measurement*, 1956, **16**, 13–37.

DE FINETTI, B. Methods for discriminating levels of partial knowledge concerning a test item. *British Journal of Mathematical and Statistical Psychology*, 1965, **18**, 87–123.

DRESSEL, P. L., & SCHMID, P. Some modifications of the multiple-choice item. *Educational and Psychological Measurement*, 1953, **13**, 574–595.

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.

FRENCH, J. W., EKSTROM, R. B., & PRICE, L. A. *Kit of reference tests for cognitive factors*. 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 knowledge on a multiple-choice test. *Journal of Educational Measurement*, 1970, **7**, 75–82.

HOPKINS, K. D., HAKSTIAN, A. R., & HOPKINS, B. R. Validity and reliability consequences of confidence weighting. *Educational and Psychological Measurement*, 1973, **33**, 135–141.

KANSUP, W., & HAKSTIAN, A. R. A comparison of several methods of assessing partial knowledge in multiple-choice tests: I. Scoring procedures. *Journal of Educational Measurement*, 1975, **12**, 219–230.

KOEHLER, R. A. A comparison of the validities of conventional choice testing and various confidence marking procedures. *Journal of Educational Measurement*, 1971, **8**, 297–303.

LORD, F. M. Formula scoring and number-right scoring. *Journal of Educational Measurement*, 1975, **12**, 7–11.

MICHAEL, J. C. The reliability of a multiple-choice examination under various test-taking instructions. *Journal of Educational Measurement*, 1968, **5**, 307–314.

RIPPEY, R. M. Probabilistic testing. *Journal of Educational Measurement*, 1968, **5**, 211–215.

RIPPEY, R. M. A comparison of five different scoring functions for confidence tests. *Journal of Educational Measurement*, 1970, **7**, 165–170.

SHUFORD, E. H., ALBERT, A., & MASSENGILL, H. E. Admissable probability measurement procedures. *Psychometrika*, 1966, **31**, 125–145.

WANG, M. W., & STANLEY, J. C. Differential weighting: A review of methods and empirical studies. *Review of Educational Research*, 1970, **40**, 663–705.

WILLIAMS, R. E., & MILLMAN, J. Confidence testing—a reply. *NCME Measurement News*, 1970, **13**, 8–9.

AUTHORS

HAKSTIAN, A. RALPH *Address:* Department of Psychology, University of British Columbia, Vancouver, British Columbia, Canada. *Title:* Associate Professor of Psychology. *Degrees:* B.A. University of British Columbia, M.A., Ph.D. University of Colorado. *Specialization:* Psychometrics; Statistics; Individual Differences.

KANSUP, WANLOP *Address:* Department of Educational Psychology, University of Alberta, Edmonton, Alberta, Canada. *Title:* Ph.D. Student. *Degrees:* B.Ed., M.Ed. The College of Education, Bangkok, Thailand; M.Ed. University of Alberta. *Specialization:* Psychometrics; Assessment; Research Methodology.