DOCUMENT RESUME

ED 090 312		TM 003 601
AUTHOR	Jacobs, Stanley S.	
TITLE	Behavior on Objective Tests Under Theoretically Adequate, Inadequate and Unspecified Scoring Rules.	
PUB DATE	74	
NOTE	18p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, Illinois, April, 1974)	
EDRS PRICE	MF-$0.75 HC-$1.50 PLUS POSTAGE	
DESCRIPTORS	*Confidence Testing; Criterion Referenced Tests; Guessing (Tests); *Multiple Choice Tests; *Risk; *Scoring Formulas; Test Interpretation; Test Validity	

ABSTRACT

Investigated were the effects of two levels of penalty for incorrect responses on two dependent variables (a measure of risk-taking or confidence, based on nonsense items, and the number of response-attempts to legitimate items) for three treatment groups in a 2x3, multi-response repeated measures, multivariate ANOVA (Analysis of Variance) design. Subjects responded under one of three scoring-administrative rules: conventional Coombs-type directions and two variants suggested as mathematically more adequate. Results indicated significant differences both among groups and across conditions. The results were discussed with reference to the question of test validity in general, and the problems posed for criterion-referenced measurement. (Author)

Division D/ NCME
18.11

Behavior on Objective Tests Under Theoretically Adequate, Inadequate and Unspecified Scoring Rules

Stanley S. Jacobs
University of Pittsburgh

Paper presented at the annual meeting of the
American Educational Research Association
Chicago, Illinois

April 1974

Behavior on Objective Tests Under Theoretically Adequate, Inadequate and Unspecified Scoring Rules

Stanley S. Jacobs, University of Pittsburgh

Abstract

Investigated were the effects of two levels of penalty for incorrect responses on two dependent variables (a measure of risk-taking or confidence, based on nonsense items, and the number of response-attempts to legitimate items) for three treatment groups in a 2 x 3, multi-response repeated measures, multivariate ANOVA design. Ss responded under one of three scoring-administrative rules: conventional Coombs-type directions and two variants suggested as mathematically more adequate. Results indicated significant differences both among groups and across conditions. The results were discussed with reference to the question of test validity in general, and the problems posed for criterion-referenced measurement.

A number of alternative administrative and scoring procedures for objective tests have been suggested (e.g. Coombs, 1953; de Finetti, 1965; Ebel, 1965; Rippey, 1968) which have as their common objective a more adequate assessment of the degree of partial knowledge held by a given student with reference to a given item.[1]

A procedure known as 'option-elimination' or 'Coombs-type directions' (CTD) seems quite applicable to the typical classroom testing situation. With CTD, the student is required to identify as many of the j-1 distractors among the j item options as he or she is able. With the usual scoring rule, a student earns one point for each distractor so identified. A penalty of -(j-1) points is suffered if the correct answer is identified as a distractor. Item scores, then, can range from -(j-1) points to +(j-1) points, having 2(j-1)+1

[1] See Echternacht (1972) for a comprehensive description and review of a number of alternative testing procedures.

possible values, rather than simply the I or 0 earned under conventional conditions. Apparently, CTD have the potential for discerning intermediate levels of knowledge.

Hritz and Jacobs (1970) demonstrated, however, that the problems associated with the correction-for-guessing, documented by Votaw (1936) and Sherriffs and Boomer (1954), have apparently been simply shifted from the item level to the option level under CTD. Using CTD all Ss behaved conservatively, identifying too few distractors. There seem to be reliable and extreme individual differences in the tendency to respond to items under an announced guessing penalty and, similarly, to identify distractors under a procedure which effectively incorporates such a penalty. Furthermore, these differential response tendencies seem to be moderated by personality variables unrelated to the variable measured by the test under scrutiny (Slakter, 1968; Jacobs, 1971).

It has been suggested that the differential tendency to "take risks" in the identification of some number of the j-I options may be controlled by increasing the level of penalty imposed for incorrect responses (Arnold and Arnold, 1970, p. 13). There is, to the author's knowledge, no direct empirical evidence for this suggestion where CTD are concerned. However, if one extends the "argument by analogy" from the similar results obtained in the Sherriffs and Boomer (1954) and Hritz and Jacobs (1970) studies, the results of Waters (1967) may be relevant. Waters found that increased levels of penalty resulted in significant increases in the number of omitted items in a conventional multiple-choice testing situation. One might hypothesize that increasing the level of penalty under CTD would result in analogous behavior; i.e., Ss will identify fewer of the j-I distractors.

Arnold and Arnold (1970) have also suggested that the usual credit-penalty arrangement used with CTD, described above, is mathematically inadequate

They then present, under a game-theoretic model, what is proposed as a more adequate system. The credit-penalty arrangement is such that the following are the "fair scores" assigned to various responses to a four-option multiple-choice item.

TABLE 1

Fair Scores for a Four-Option Multiple-Choice Item, as Developed by Arnold and Arnold (1970), and Used in the "A & A, Specified" Condition in the Present Study

Outcome	Fair Score
Including the correct answer in the set of options identified as distractors	-1
No distractors identified	0
One distractor correctly identified	1/3
Two distractors correctly identified	1
Three distractors correctly identified	3

Unfortunately, some of the derivations involve the introduction of the assumption of a random-guessing model, similar to that which is involved in the derivation of the usual guessing correction formulae. It seems inconsistent to develop a model for behavior under CTD (which theoretically involves a rational partitioning of item options into two sets, one of which S feels contains the correct answer) which assumes a random process in responding. Also, the data offered by Arnold and Arnold in support of their scoring procedure are suspect, since Ss in their study were simply told that if they guessed their expected gain would be zero. No actual information as to credit or penalty was provided to give Ss some basis for decision making. There is evidence that even minor changes in the directions provided Ss can produce significant changes in test scores, on measures of cognitive variables (Yamamoto and Dizney, 1965) and on measures of personality (Jacobs, 1972), obtained under more conventional conditions.

The purpose of the present study was to examine the effects of increasing penalty level, under three types of CTD instructions, on performance on a multiple-choice vocabulary test. The three credit-penalty arrangements were: (1) the conventional approach described above and (2) two variants of the Arnold and Arnold approach; (a) one with all weights specified and (b) one without, with Ss simply informed that guessing would result in a zero expected gain (or, in the case of increased penalty, an expected loss to S). Under 1 and 2a, announced penalties were doubled under the increased penalty condition.

Method

Materials

Two randomly parallel 50 item multiple-choice vocabulary tests, which included 10 nonsense items for use with Slakter's (1967) measure of risk-taking were developed for the study. (See Table 2 for descriptive data.)

TABLE 2

Descriptive Data for the Two Randomly Parallel Vocabulary Tests Used in the Present Study

Form	n	k	$\bar{x}$	s.d.	t*	F**	r_{AB}
A	32	40	24.9	5.3	.99 (ns)	1.59 (ns)	.76
B	32	40	25.6	6.7			

*test on means
**test on variances

The data presented in Table 2 are based on the 40 legitimate items only, and were collected on a group of 32 Ss similar to those in the present study. The tests were administered in a constant A-B order for all Ss, under instructions which indicated Ss should respond to all items; scores were based on the number correct. The t and F tests indicate that the means and variances, respectively, are not significantly different (p > .05). The between-forms correlation of .76 indicates the two forms produce data sufficiently consistent to justify their use in a repeated measures experiment.

Two dependent variables were used: a measure of confidence or risk-taking, based on Slakter's (1967) formula and responses to the 10 nonsense items on each form, and a related index, the number of item options responded to on the 40 legitimate items on each form. The differences in scoring rules makes a comparison of actual scores meaningless if computed according to the rules presented.

Subjects

After development and tryout of the two vocabulary tests to be used in the present study, subjects were recruited from the enrollment of two sections of the introductory master's level research methods course in the School of Education at the University of Pittsburgh. They were asked to participate in a study to be conducted during an hour of class time, which would serve as a vehicle for subsequent lectures and discussions throughout the term. All students agreed to participate.

Procedure

The 87 Ss were randomly assigned to the three treatment groups mentioned above. Each S received instructions, several worked examples, and test materials to enable individual work without need for explanation from E. This allowed Ss in different treatments to remain in the same room. An error in assignment resulted in one S being misassigned (note the n's in Table 3). Thirty minutes were allowed for form A, whereupon all materials were collected, and materials for form B distributed. Thirty minutes were also allowed for form B. All materials were than collected, and Ss were thoroughly debriefed concerning the study.

Design

The design of the study was conceptualized as a 3 x 2 full rank multi-response repeated measures design and was analyzed as such using procedures discussed by Timm and Carlson (1973) and Timm (1974).

The design of the study is presented in Figure 1, to enable the reader to relate the various hypothesis tests to the design.

Treatments	Dependent Variable 1(V_1)		Dependent Variable 2(V_2)	
	Condition 1(C_1)	Condition 2(C_2)	Condition 1(C_1)	Condition 2(C_2)
T_1	μ_{11}	μ_{12}	μ_{13}	μ_{14}
T_2	μ_{21}	μ_{22}	μ_{23}	μ_{24}
T_3	μ_{31}	μ_{32}	μ_{33}	μ_{34}

Fig. 1. Plan of the 3 x 2 Multi-response Repeated Measures Design Used in the Present Study

Results

Results descriptive of the effects of treatments and conditions on the two dependent variables employed are presented in Table 3 in a format consistent with Figure 1, and illustrated in Figures 2 and 3. The intercorrelations of the dependent variables are presented in Table 4.

The data presented in Table 3 indicate that the increased penalty had a similar impact on the two dependent variables in the three experimental groups; in all cases, increased penalty (either specified or implied) resulted in a decrease in index averages. It may also be seen that the two dependent variables are significantly and substantially correlated within penalty conditions within treatment groups.

TABLE 3

Means and Standard Deviations of the Two Dependent Variables Used, Under the Two Penalty Conditions for the Three Experimental Groups in the Present Study

	Penalty Conditions and Variables								
	risk (V_1)					response attempts (V_2)			
	low penalty(C_1)			high penalty(C_2)		low penalty(C_1)		high penalty(C_2)	
	n	$\bar{x}$	s.d.	$\bar{x}$	s.d.	$\bar{x}$	s.d.	$\bar{x}$	s.d.
D(T_1)	30	14.60	6.74	6.97	7.96	81.73	14.20	71.70	15.71
A, specified(T_2)	28	17.68	8.49	10.18	8.30	85.82	17.56	76.39	20.72
A, unspecified(T_3)	29	19.59	6.72	12.59	7.72	93.55	15.62	84.86	17.18

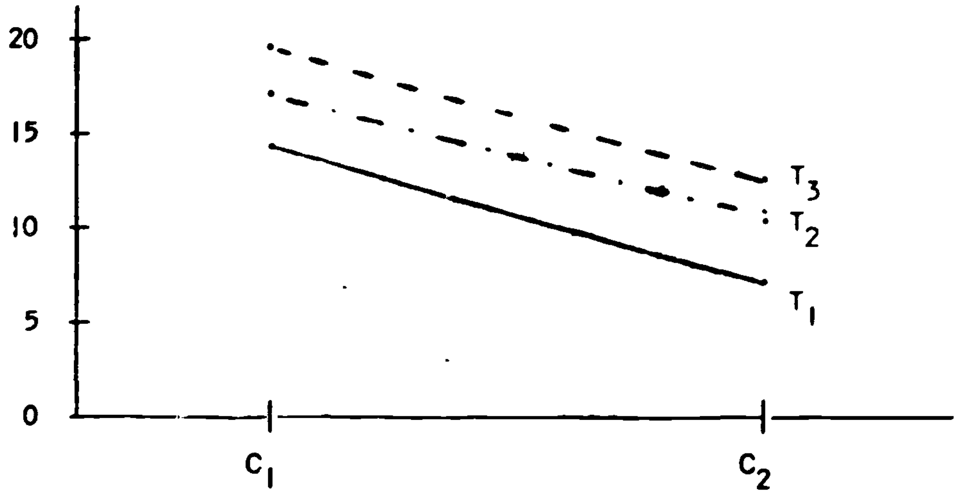

Fig. 2. Flot of Data Points for Variable One, Risk-taking or Confidence, Across Two Penalty Conditions (C_1 and C_2) for the Three Treatment Groups

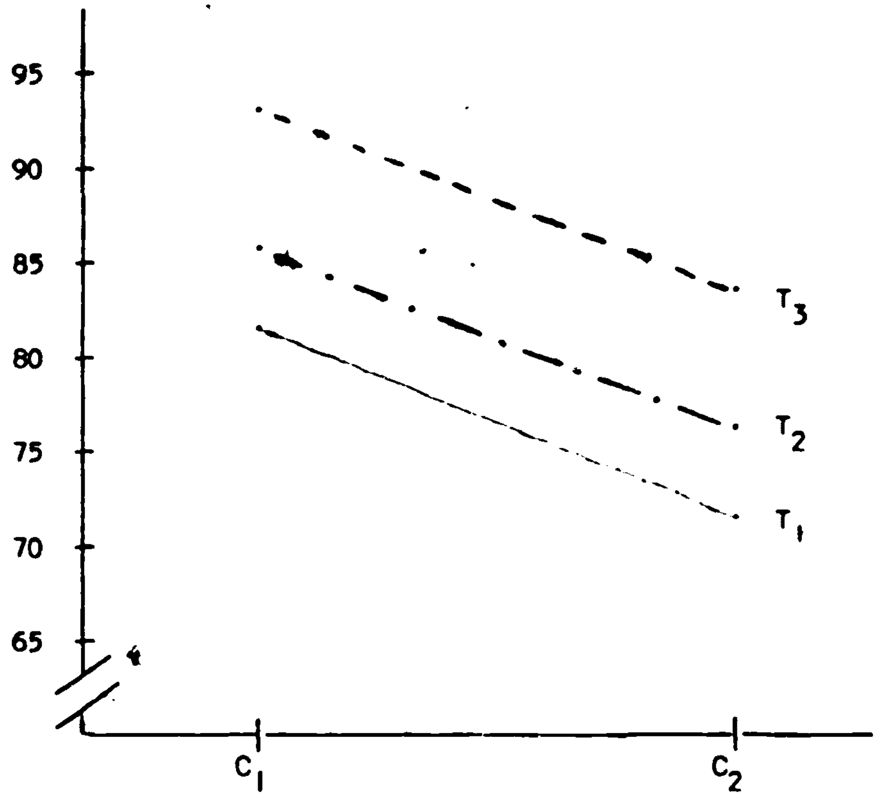

Fig. 3. Plot of Data Points for Variable Two, Number of Response Attempts, Across the Penalty Conditions (C_1 and C_2), for the Three Treatment Groups

A multivariate F-test was performed to test differences among treatments;

$$H_o: \begin{pmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{13} \\ \mu_{14} \end{pmatrix} = \begin{pmatrix} \mu_{21} \\ \mu_{22} \\ \mu_{23} \\ \mu_{24} \end{pmatrix} = \begin{pmatrix} \mu_{31} \\ \mu_{32} \\ \mu_{33} \\ \mu_{34} \end{pmatrix}$$

In other words, the significance of differences among the T_1, T_2 and T_3 mean vectors was tested. The results of this analysis are presented in Table 5; the multivariate $F_{8,162}$ of 1.795 has a chance probability of approximateiy .08. (See Table 5)

TABLE 4

Intercorrelations of the Two Dependent Variables Under the Two Penalty Conditions for the Three Experimental Groups of the Present Study[1]

Experimental Groups		Treatment Conditions and Variables			
		risk (V_1)		response attempts (V_2)	
		low penalty(C_1)	high penalty(C_2)	low penalty(C_1)	high penalty (C_2)
	V_1C_1	1.00			
CTD (T_1) (n = 30)	V_1C_2	.63	1.00		
	V_2C_1	.85	.29	1.00	
	V_2C_2	.49	.55	.52	1.00
	V_1C_1	1.00			
A & A pecified(T_2) (n = 28)	V_1C_2	.68	1.00		
	V_2C_1	.88	.55	1.00	
	V_2C_2	.55	.79	.65	1.00
	V_1C_1	1.00			
A & A nspecified(T_3) (n = 29)	V_1C_2	.45	1.00		
	V_2C_1	.85	.22	1.00	
	V_2C_2	.45	.58	.54	1.00

[1] for n = 30, df = 28, $r > .306$, $p < .05$ and $r > .463$, $p < .01$
for n = 29, df = 27, $r > .311$, $p < .05$ and $r > .471$, $p < .01$
for n = 28, df = 26, $r > .317$, $p < .05$ and $r > .479$, $p < .01$

TABLE 5

Multivariate ANOVA Summary Table; "Treatments" Hypothesis

Source	DF	SSP	Multivariate F	p-value
reatments (T)	8	$\begin{pmatrix} 382.112 & & & \text{(Sym)} \\ 801.143 & 2115.842 & & \\ 392.378 & 964.627 & 469.501 & \\ 893.491 & 2351.827 & 1075.491 & 2614.493 \end{pmatrix}$	1.795	.0814
rror	162	$\begin{pmatrix} 4311.842 & & & \text{(Sym)} \\ 8185.214 & 21005.146 & & \\ 2861.542 & 5839.247 & 5366.108 & \\ 5438.981 & 13775.771 & 7836.081 & 27008.427 \end{pmatrix}$		

This analysis tests for the coincidence of the four data points obtained within each of the three treatments. The overall F is regarded as indicating a significant departure from coincidence ($p < .10$).

Confidence intervals were calculated for differences between treatments' data points. It was determined that significant differences for V_1C_1, V_2C_1, V_2C_1, and V_2C_2 existed between T_1(CTD) and T_3 (A & A, unspecified), only ($p < .10$).

A multivariate F-test was then performed to test the significance of differences across the two penalty conditions, C_1 and C_2, simultaneously for both variables,

$$H_o: \begin{pmatrix} \mu_{11} \\ \mu_{21} \\ \mu_{31} \\ \mu_{13} \\ \mu_{23} \\ \mu_{33} \end{pmatrix} = \begin{pmatrix} \mu_{12} \\ \mu_{22} \\ \mu_{23} \\ \mu_{14} \\ \mu_{24} \\ \mu_{34} \end{pmatrix}$$

The multivariate $F_{6,166}$ of 14.702 has a chance probability of less than .001; a summary of the analysis is presented in Table 6.

TABLE 6

Multivariate ANOVA Summary Table; "Conditions" Hypothesis

Source	df	SSP	Multivariate F	p-value
Conditions (C)	6	$\begin{pmatrix} 4884.134 & 6131.934 \\ 6131.934 & 7698.971 \end{pmatrix}$	14.702	< .001
Error	166	$\begin{pmatrix} 3954.867 & 6743.067 \\ 6743.067 & 20462.031 \end{pmatrix}$		

The calculation of confidence intervals for the obtained differences across penalty conditions indicated that the increased level of penalty had a significant effect ($p < .05$) on variable 1 (a measure of confidence or risk-taking), but not variable 2 (the number of responses made to the options of legitimate items), and the effect was similar for all three treatment groups.

A multivariate F-test of the interaction (T x C) hypothesis produced a nonsignificant $F_{4,166}$ of 0.70 ($p = .9911$). In view of the nonsignificant interaction, two additional analyses were performed.

The first analysis compared differences among treatment groups by contrasting vectors of V_1, V_2 averages simultaneously, collapsed across conditions:

$$H_0: \begin{pmatrix} \frac{\mu_{11}+\mu_{12}}{2} \\ \\ \frac{\mu_{13}+\mu_{14}}{2} \end{pmatrix} = \begin{pmatrix} \frac{\mu_{21}+\mu_{22}}{2} \\ \\ \frac{\mu_{23}+\mu_{24}}{2} \end{pmatrix} = \begin{pmatrix} \frac{\mu_{31}+\mu_{32}}{2} \\ \\ \frac{\mu_{33}+\mu_{34}}{2} \end{pmatrix}$$

The multivariate F-test produced an $F_{4,166}$ of 2.941, which has a chance probability of approximately .0221. The results of this analysis are summarized in Table 7.

TABLE 7

Multivariate ANOVA Summary Table; "Treatments'" Hypothesis

Source	df	SSP	Multivariate F	p-value
Treatments' (T')	4	$\begin{pmatrix} 395.592 & 933.689 \\ 933.689 & 2358.497 \end{pmatrix}$	2.941	.0221
Error	166	$\begin{pmatrix} 3850.258 & 6324.880 \\ 6324.880 & 18891.279 \end{pmatrix}$		

Confidence intervals calculated indicated the significant differences were between T_1 (CTD) and T_3 (A&A, unspecified), only ($p < .05$) and only for variable 1 (risk-taking, collapsed across conditions).

The second analysis contrasted conditions C_1 and C_2 by contrasting vectors of V_1, V_2 averages simultaneously over the three treatment groups:

$$H_o: \begin{pmatrix} \sum_i \mu_{11}/3 \\ \\ \sum_i \mu_{13}/3 \end{pmatrix} = \begin{pmatrix} \sum_i \mu_{12}/3 \\ \\ \sum_i \mu_{14}/3 \end{pmatrix}$$

The multivariate F-test resulted in an $F_{2,83}$ of 55.566, $p < .0001$. A summary of this analysis is presented in Table 8.

TABLE 8

Multivariate ANOVA Summary Table; "Conditions'" Hypothesis

Source	df		Multivariate F	p-value
Conditions' (C')	2	$\begin{pmatrix} 4860.935 & 6099.982 \\ 6099.982 & 7654.860 \end{pmatrix}$	55.566	< .0001
Error	83	$\begin{pmatrix} 3954.867 & 6743.067 \\ 6743.067 & 20462.031 \end{pmatrix}$		

When confidence intervals were calculated for obtained differences across conditions C_1 and C_2 for variables (V_1 and V_2) collapsed across treatments (T_1, T_2 and T_3), it was found that differences for both variables were significant. That is, when "condition effects" are obtained by averaging across treatment groups, the effect is significant for both dependent variables ($p < .05$).

Summary and Discussion

Although the descriptive data presented in Table 3 and in Figures 2 and 3 seem to indicate that increased penalty-level has a similar effect on both dependent variables, several analyses indicate that V_1, Slakter's measure of risk-taking, is apparently the more sensitive and consistent dependent variable.

The test of the first "treatments" hypothesis (see Table 5) indicated significant differences between T_1 and T_3 for both dependent variables and both conditions, but with $p < .10$. A second test of this hypothesis, based on V_1, V_2 averages, collapsed across conditions indicated that significant differences existed between T_1 and T_3 ($p < .05$), but only for V_1 (see Table 7).

When both "conditions" hypotheses were tested (see Tables 6 and 8), V_1 reflected a significant decline from C_1 to C_2 in both analyses; V_2 did so only in the latter analysis.

Variable one, then, a measure of risk-taking (as Slakter termed it) or, more descriptively, a measure of confidence exhibited in attempting to answer what probably appear to Ss as very difficult items, seems to be more affected by both the increase in penalty and the implications in test directions.

Although it is usually not considered good practice to develop procedures under one set of conditions, with the expectation they will readily generalize to another set, the data of the present study indicate that the differences in behavior between conditions where the details of a scoring rule are presented, and where they are not, are not statistically significant. However, with respect to CTD in their usual form, and to a condition where penalties are implied but not specified, it appears that students exhibit more confidence in attempting to answer nonsense items (which, at least logically, may have a "stimulus-value" analogous to an extraordinarily difficult legitimate item) under the latter conditions. These results are consistent with those of Waters, who found that students apparently view completely unspecified scoring weights as indicating zero weights for incorrect answers.

The results of the present study indicate that the problem of conservative responding under CTD noted by Hritz and Jacobs (i.e. all Ss tended to identify too few distractors) may be partially resolved using procedures paralleling those of T_3. The question of the effect on test validity and the possible interaction with subject (attribute) variables needs investigation. Also, one is left wondering what the long term effects of experience with the procedures in the present study would be, e.g. after some experience, would Ss in T_2 and T_3 condition behave in a more similar fashion? Would an interaction between level of penalty and subject characteristics develop?

The present study has implications for the general domain known as criterion-referenced testing. While most of the effort in this area thus far has centered around strategies for item and test development, item and test administration and item and test analysis, a very fundamental question remains; what should Ss be told when they confront a criterion-referenced test? The results of the present study indicate that behavior observed may vary as a function of item difficulty, instructions to the S, and penalty for incorrect responses.

Since behavior is compared with some criterion or standard, one cannot assume that the effect (if any) is constant across all Ss (which would permit legitimate norm-referenced comparisons) therefore of no importance. One may instead consistently over-or under-estimate the level of performance of a group of Ss, depending upon how the criterion information was generated.

References

Arnold, J.C. and Arnold, P.L. On scoring multiple choice exams allowing for partial knowledge. *The Journal of Experimental Education*, 1970, 39, 8-13.

Coombs, C.H. On the use of objective examinations. *Educational and Psychological Measurement*, 1953, 13, 308-310.

de Finetti, B. Methods for discriminating levels of partial knowledge concerning a test item. *The British Journal of Mathematical and Statistical Psychology*, 1965, 18, 87-123.

Ebel, R.L. Confidence weighting and test reliability. *Journal of Educational Measurement*, 1965, 11, 49-57.

Echternacht, F.J. The use of confidence testing in objective tests. *Review of Educational Research*, 1972, 42, 217-236.

Hritz, R.J. and Jacobs, S.S. Risk-taking and the assessment of partial knowledge. Paper presented at the annual meeting of the American Psychological Association, Miami Beach, Florida, September 1970.

Jacobs, S.S. Correlates of unwarranted confidence in responses to objective test items. *Journal of Educational Measurement*, 1971, 8, 15-20.

Jacobs, S.S. A validity study of the acquiescence scale of the Holland Vocational Preference Inventory. *Educational and Psychological Measurement*, 1972, 32, 477-480.

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

Sherriffs, A.C. and Boomer, D.S. Who is penalized by the penalty for guessing? *Journal of Educational Psychology*, 1954, 45, 81-90.

Slakter, M.J. Risk-taking on objective examinations. *American Educational Research Journal*, 1967, 4, 31-43.

Slakter, M.J. The penalty for not guessing. *Journal of Educational Measurement*, 1968, 5, 141-144.

Timm, N.H. Multivariate profile analysis of split-split plot designs and growth curve analysis of multivariate repeated measures designs. Paper presented at the annual meeting of the American Educational Research Association, Chicago, Illinois, April 1974.

Timm, N.H. and Carlson, J.E. Multivariate analysis of nonorthogonal experimental designs using a multivariate full rank model. Paper presented at the annual meeting of the American Statistical Association, New York, New York, December 1973.

Votaw, D.F. The effect of do-not-guess directions upon the validity of true-false or multiple-choice tests. *Journal of Educational Psychology*, 1936, 28, 698-702.

Waters, L.K. Effect of perceived scoring formula on some aspects of test performance. *Educational and Psychological Measurement*, 1967, 27, 1005-1010.

Yamamoto, K. and Dizney, H.F. Effects of three sets of test instructions on scores on an intelligence scale. *Educational and Psychological Measurement*, 1965, 25, 87-94.