On the Feasibility of Multiple Matching Tests—Variations on a Theme by Gulliksen

David V. Budescu
University of Haifa

This paper reports a feasibility study of a new test format—multiple matching (MM). Under this format, distractors from several items are pooled into a single list which appears at the end of the test. The examinees are asked to match one correct answer to each of the items in the test. An experiment is described in which the lengths of the item list and the distractor pool were manipulated. It is shown that construction of MM tests is technically feasible and that these tests achieve satisfactory results in terms of reliability, validity, and reduction of random guessing.

It is widely recognized that persons taking multiple-choice tests can select correct answers to items that they could not have answered in an open-ended format. This fact is usually referred to as "the guessing problem" and has been the subject of a large body of theoretical and empirical research (for partial reviews of this work, see, e.g., Abu-Sayf, 1977; Diamond & Evans, 1973; Hutchinson, 1982). Most studies have focused on scoring formulas or rules, designed to eliminate the effect of (or penalize for) random guessing—and, in some cases, to reward partial knowledge. The best known and most popular formula (e.g., Lord, 1975) is

$$S_1 = R - \frac{W}{a-1} \quad , \tag{1}$$

where R is the number of correct answers,
W is the number of incorrect answers, and
a is the number of options per item.

Additional scoring procedures have been proposed by Abu-Sayf (1977), Reilly (1975), Traub, Hambleton, and Singh (1969), and Zinger (1972) within the framework of classical test theory, and by van der Ven (1974) and Molenaar (1977) in a Bayesian context. In item response theory, the three-parameter model explicitly incorporates a "pseudo-guessing" parameter which affects the ability estimates (e.g., Lord, 1980), and Wainer (1983) recently proposed robust scoring rules which are less sensitive to unusual responses (some of which may be attributed to guessing).

The key to the success of any scoring rule is its capability to distinguish between various levels of partial knowledge, and to reward (or penalize) the examinees accordingly. Thus it is extremely important that all examinees understand the nature of the rule and all its implications, including the optimal response strategy when they are uncertain of the correct answer. Furthermore, examinees must be willing to follow instructions.

Empirical evidence suggests that these conditions cannot be realistically expected. Cross and Frary (1977) and Frary and Hutchinson (1982) showed that even highly motivated examinees do not always understand, remember, or follow the instructions of the simplest scoring rule (Equation 1), and thus do not obtain full credit for their true

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 12, No. 1, March 1988, pp. 5–14

0146-6216/88/010005-10$1.75

level of knowledge. In addition, several studies have linked the tendency to ignore these instructions with specific personality traits (e.g., Sherriffs & Boomer, 1954; Slakter, 1969; Swineford, 1941; Ziller, 1957).

A correction rule is obviously invalid if examinees do not follow it. Instead of minimizing the effect of irrelevant variance, it may actually add new sources of measurement error. Thus, the problem of guessing is not a purely psychometric issue. It is, to a large degree, a problem of design. The success of any test developer in approaching it should be evaluated by his/her ability to design items, instructions, and a scoring rule that work well together. In other words, a test developer should be able to devise a scoring rule, and appropriate instructions, that will be understood, accepted, and followed by most (if not all) examinees in the context of a specific examination.

Other solutions to the guessing problem rely on alternative item formats (e.g., Carlson, 1985; Gulliksen, 1986) rather than special scoring rules. One of Gulliksen's proposals that has received special notice (Wainer, 1983; Wood, 1984) because of its intuitive appeal is the multiple matching (MM) format. The suggestion is to replace the stem-plus-options format by a list of items followed by a longer list of options. Note that if the number of items equals the number of options, the test becomes a simple matching (SM) task (e.g., Horst, 1966; Wesman, 1971). Such a design is easy to implement; item writing is straightforward, the format discourages item omission, and the probability of guessing the correct answers is generally reduced. The scoring rule is also simple: S_2 = number of correct matches. Zimmerman and Williams (1982) showed that in SM, when examinees either know the answer or guess at random, the expected score is

$$E(S_2/K) = K + 1 \quad . \tag{2}$$

This can be compared with the expected score under a regular multiple-choice (MC) format:

$$E(S_1/K) = K + \frac{n-K}{a} \quad , \tag{3}$$

where n is the total number of items and K is the number of items answered correctly.

The SM rule yields a fixed chance component for each level of K, while the magnitude of this component in MC tests is a linear function of K. The two scores coincide when $K = (n - a)$. However, the chance component of S_1 is higher for all examinees who know less than $(n - a)$ answers, and lower for those who can answer more than $(n - a)$ items correctly. Given that, in most cases, a is quite small in comparison with n, it can be inferred that the large majority of examinees will benefit more from guessing under the MC format. Yet a minority of high-ability examinees (i.e., those who need to randomly match fewer than a items) will score higher under a SM paradigm.

Under MM, all na options of a regular MC test can be pooled in order to create a (very) long list of options from which the examinee must select n answers. If S_3 is defined as the number of correct matches, then it is easy to show that the expected score of an examinee who knows K answers is

$$E(S_3/K) = K + \frac{n-K}{na-K} \tag{4}$$

(see the Appendix), and that the chance component is considerably smaller than its counterparts in S_1 and S_2. The relative magnitude of the chance component is monotonically (but nonlinearly) decreasing as a function of both K and a. In other words, the impact of purely random guessing is greater for low-ability examinees and can be decreased by increasing the number of options. Finally, note that SM is just a special case of this model, which is obtained by setting $a = 1$.

The purpose of the present study was to compare the two forms along several dimensions in order to examine the feasibility of the MM format for practical use. Research comparing MC and SM matching formats is scarce, and there has been no research on MM items. As mentioned earlier, Zimmerman and Williams (1982) have shown theoretically that SM tests usually have higher reliability, and Baldauf and Propst (1979) and Baldauf (1982) have reinforced this conclusion with empirical data and monte carlo simulations.

Method

A vocabulary test of 24 items was constructed from the item pool of the National Institute of Testing and Evaluation (NITE). The test consisted of three subtests of equal length: Hebrew nouns, Hebrew verbs, and foreign words used in Hebrew (such as sarcophagus, choral, etc.). Thus, three subtests with relatively homogeneous items were obtained. Originally, the items were administered in MC format with $a = 4$ options. For the present study, six additional MM versions were prepared. The six versions represent a 3×2 design in two independent variables: a, the number of options per item, and n, the number of items. Three values of a (1, 2, and 4) and two values of n (4 and 8) were used. For the manipulation of n each 8-item subtest was randomly halved, and all the relevant options of that half test were pooled. For the manipulation of the number of options the correct answer was used alone ($a = 1$), in conjunction with the most attractive distractor ($a = 2$), or with all the original distractors ($a = 4$). In each of the six MM forms the n items were listed at the top of the page followed by the na distractors listed in alphabetical order.

The examinee pool consisted of 717 applicants to various universities in Israel who took the national admission test in August 1985 or April 1986. The test is similar to the Scholastic Aptitude Test (SAT) and its administration requires about 2½ hours.

Procedure

The six experimental test versions and the regular MC version were administered in a random sample of testing classrooms after the completion of the admission test. The special matching format was explained and illustrated by examples. The instructions stressed that no penalty for guessing was used, and that it was in the examinee's interest to answer all items. All examinees completed the test in less than 30 minutes, and their completion times were recorded. After the test, all examinees in the MM groups completed (anonymously) a short feedback questionnaire comparing the experimental forms to the regular MC. Table 1 summarizes the design and sample sizes in each condition. Note that the $a = 4, n = 1$ cell represents the MC format.

Results

Table 2 presents the means and the standard deviations of the three tests (and their aggregated total) under each of the seven administration modes. Although this table summarizes the data as a function of a and n, it is important to keep in mind that an equally good representation of the results can be obtained by pooling cells with an equal number of options, $T = na$. Two conditions ($a = 4$, $n = 1$ and $a = 1$, $n = 4$) have a total of $T = 4$ options, and two conditions ($a = 2$, $n = 4$ and $a = 1$, $n = 8$) have $T = 8$ options; two cells ($a = 4$, $n = 4$ and $a = 2$, $n = 8$) have $T = 16$ options, and one cell ($a = 4$, $n = 8$) has $T = 32$ total options. Examination of Table 2 indicates that, as expected, the mean score decreases as a and/or n increase. Within each of the table's rows and columns (with two minor exceptions) the mean score decreases, and a comparison of forms with equal numbers of distractors indicates that the effect of a is stronger in this context. Analysis of variance (ANOVA) of the results showed a significant effect for number of options ($F_{2,710} = 114.88$, $p<.05$) and number of items ($F_{2,710} = 36.81$, $p<.05$), but no significant interaction ($F_{2,710} = 2.89$, $p>.05$).

The pattern of change in the scatter of the scores is not as clear and consistent as far as a is concerned. However, note in Table 2 that for each

Table 1
Sample Sizes for Testing Conditions Defined by the Number of Items Pooled and the Number of Options Per Item

Number of Items Pooled	Number of Options Per Item			Total
	1	2	4	
1	-	-	110	110
4	72	107	111	290
8	75	123	119	317
Total	147	230	340	717

Table 2
Means and Standard Deviations of Scores for the Subtests Under the Seven Administration Forms

Number of Items Pooled and Subtest	Number of Options Per Item					
	1		2		4	
	Mean	SD	Mean	SD	Mean	SD
1 Item						
N	-		-		4.06	1.99
V	-		-		4.56	1.74
F	-		-		4.93	1.48
Total	-		-		13.55	3.90
4 Items						
N	5.86	2.12	2.91	1.51	2.88	2.31
V	6.71	1.77	4.81	1.44	3.64	1.80
F	6.76	1.92	5.29	1.72	3.88	1.69
Total	19.33	4.60	13.01	3.52	10.41	4.85
8 Items						
N	4.56	2.51	3.90	2.31	2.75	2.31
V	5.84	1.83	4.56	1.75	3.18	1.84
F	5.76	2.08	4.78	1.93	3.04	1.84
Total	16.16	5.18	13.24	4.85	8.97	4.91

subtest, the standard deviation increases as n increases.

Table 3 presents the same data as a function of T, ignoring the distinction between a and n. It is clear that as T increases, the mean scores decrease and the variances increase.

Table 4 summarizes five measures of inter-item homogeneity as a function of T, the total number of alternatives. The statistics are the KR-20 reliability, the median item-total biserial correlation, the mean inter-item correlation, the mean inter-subtest correlation, and the proportion of variance accounted for by the first principal component of the inter-item correlation matrix. Although the differences are not large, they all indicate that an increase in the number of options yields more homogeneous, reliable, and unidimensional tests.

Obviously, these improvements are achieved at a certain cost. In this case, one component of the cost is the total testing time. As Table 5 shows, this time increases as a function of a and n. When data are summarized according to the total number

Table 3
Means and Pooled Standard Deviations of Scores as a Function of the Total Number of Alternatives

Total Number of Alternatives	Subtest							
	N		V		F		Total	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
4	4.77	2.04	5.41	1.75	5.65	1.67	15.84	4.19
8	3.59	1.98	5.23	1.61	5.48	1.88	14.31	4.28
16	3.42	2.31	4.12	1.77	4.35	1.82	11.90	4.85
32	2.75	2.31	3.18	1.84	3.04	1.84	8.97	4.91

Table 4
Inter-Item Homogeneity as a Function of Total Number of Alternatives

Number of Alternatives	KR20	Median Biserial	Mean Inter-Item Correlation	Mean Inter-Test Correlation	Percent Variance of Principal Component
4	.77	.39	.12	.57	12
8	.77	.42	.15	.58	15
16	.82	.46	.16	.68	16
32	.84	.46	.16	.70	16

of alternatives, a similar pattern emerges. The mean testing times are 6.16, 6.89, 9.49, and 12.98 minutes for $T = 4$, 8, 16, and 32, respectively. An ANOVA of the rates of response (i.e., the reciprocals of the total testing times) reveals highly significant effects of a ($F_{2,684} = 127.7$, $p<.05$) and n ($F_{2,684} = 102.58$, $p<.05$) as well as a significant interaction ($F_{2,684} = 5.09$, $p<.05$).

Table 6 displays the distribution of the incorrect responses for those forms where examinees could select either original or new distractors ($a>1$). Note that the omission rate is negligible and equal under all forms. In all forms, an absolute majority of the incorrect responses involves new distractors (i.e., distractors that originally were written for different items). However, when these values are adjusted for the different number of distractors actually selected, it appears that the probability of choosing one of the original distractors is considerably higher. The probability of selecting one of the new distractors is always slightly smaller than its chance level (in the four cases in the table, chance levels are .167, .083, .071, and .036, respectively). Obviously, these distractors attract a large proportion of random guessing.

To examine the validity of the various MM forms, the examinees' scores were correlated with their scores on various subtests of the admission test. Because data were not available for all members of the sample, this analysis was based on only 457 examinees and does not include the results of the $a = 4$, $n = 1$ group. Table 7 reports validities with respect to two verbal subtests (Verbal General Knowledge and English), two quantitative tests (Figures and Mathematical Reasoning), and the Total Score on the admission test.

An interesting pattern emerges: The validity of the experimental tests with respect to the verbal subtests and the total score increases monotonically with T. On the other hand, the validity with respect to the quantitative tests decreases for larger values of T. The most extreme case of MM ($T = 32$) follows the same consistent pattern in comparison with the MC test: Its validity exceeds that of the regular test with respect to other verbal tests and the total scores, but is lower with respect to quantitative criteria.

What did the examinees think of the new format? The first issue is whether they understood it. The majority of the examinees (75.56%) reported that

Table 5
Mean Testing Time Under the Seven Administration Forms

Number of Items Pooled	Number of Options Per Item			Mean
	1	2	4	
1	--	--	6.77	6.77
4	5.24	7.26	10.98	8.18
8	6.36	8.15	12.98	9.54
Mean	5.81	7.74	10.32	8.56

the instructions were clear, and this figure was almost identical across all six forms.

Table 8 presents the distribution of examinee responses as a function of T, the total number of options. The assessments and preferences of the examinees are obviously related to this parameter. The SM format ($T = 4$) was judged easier overall, better promoting the examinees' interests and allowing easier guessing; in general, it was preferred by most respondents. At the other extreme, when MM is based on a long list of 32 options, most examinees found guessing, and the test, more difficult. Consequently, in this case there is a slight preference for the MC format. At the intermediate levels, responses were less extreme. For example, for $T = 8$, guessing was judged equally easy (or difficult) in both forms; a slight preference for the MC format was nonetheless evident.

Discussion

The MM method is based on a simple and attractive notion. To overcome the guessing problem, instead of using special scoring formulas based on various explicit and implicit assumptions about the response mechanism employed by the examinee, guessing is made difficult by drastically increasing the number of options available. As a matter of economical and practical convenience this can be achieved by pooling together the distractors of several independent items. If the items' content, format, and syntactical structure are comparable, such tests can be constructed.

The basic issue addressed above is whether it is feasible to construct such a test, and if its intended goals can be achieved. The results lead to a positive answer to the first question, and a somewhat qualified positive answer to the second.

With some special care it was possible to create a MM test of vocabulary. The format of the test required some special manipulations. For example, it was necessary to divide the original item pool in order to obtain three subtests with similar and homogeneous items. Furthermore, some minor adjustments and corrections were necessary in the wording of the various distractors to make each of

Table 6
Distribution of Incorrect Responses for Original and New Distractors for Tests Varying in Number of Items Pooled (n), Number of Options Per Item (a), and Total Number of Alternatives (T)

			Original Distractors					New Distractors				
n	a	T	Percent Omitted	Percent Incorrect	Percent of Total Incorrect	Total Used	Percent Used	Probability of Selecting One	Percent of Total Incorrect	Total Used	Percent Used	Probability of Selecting One
4	2	8	.01	40.31	45.16	.71	71.00	.636	54.84	4.29	71.50	.128
4	4	16	.02	55.88	46.78	1.71	57.00	.274	53.22	9.25	77.08	.057
8	2	16	.02	44.13	43.29	.96	96.00	.451	56.71	8.87	63.36	.064
8	4	32	.03	60.18	33.45	1.58	52.67	.212	66.55	16.50	51.56	.040

Table 7
Validity of MC and MM Tests as a Function of the Total Number of Alternatives (T)

Variable	MC	MM		
		T=8	T=16	T=32
Sample size	108	95	149	105
Criterion test				
Verbal General Knowledge	.72	.42	.69	.74
English as Foreign Language	.50	.49	.50	.64
Mathematical Reasoning	.51	.47	.55	.38
Figures	.33	.48	.36	.26
Total Score	.64	.60	.67	.72

them compatible with all of the items. Obviously, special attention should be given to these details in constructing MM forms. It is well known (e.g., Smith, 1982) that MC items can be solved through efficient processing of structural and linguistic clues provided by the distractors. This problem is obviously more serious in the case of the MM test, but clearly not insoluble. The unusual format calls for special instructions to the examinees, but this has caused no special problems, as became evident from the analysis of the results and the feedback questionnaires.

Assuming that it is technically feasible to construct MM tests, does it pay to do it? The method has a very obvious cost associated with it, namely longer testing time, and its efficiency should be evaluated as a function of this cost. Doubling the number of distractors inflates testing time by a factor of 1.37. If the average time required to complete a test with $T = 4$ distractors is t, then a test with

Table 8
Responses to Feedback Questionnaire, In Percentages, as a Function of the Total Number of Alternatives

Question	Total Number of Alternatives			
	4	8	16	32
Format Preferred				
MC	22.22	48.26	51.56	45.09
MM	61.11	34.14	33.33	41.18
Indifferent	16.67	17.60	15.11	13.73
Guessing is easier				
Under MC	16.67	26.40	44.89	50.00
Under MM	48.61	23.11	19.11	16.67
Same	34.72	50.49	36.00	33.33
Examinees' interests are better served				
Under MC	29.85	59.27	64.45	38.82
Under MM	70.15	40.73	35.55	41.18
Test is easier				
MC	39.71	51.38	62.17	69.70
MM	60.29	48.62	37.83	30.30

$T = 8$ requires on the average $1.12t$, one with $T = 16$ requires $1.54t$, and a test with 32 distractors requires $2.11t$. Note that although the increase in cost when moving from $T = 4$ to $T = 8$ is modest, further lengthening the test is more expensive.

At least two statistics, reliability and validity, lend themselves to relatively simple cost/benefit analyses. Classical test theory long ago established the expected relationships between test length and these statistics in cases where the length is altered by adding or removing items (e.g., Horst, 1966; Lord & Novick, 1968). Using the KR-20 of the test with $t = 1$ as a baseline it is possible to predict, through the Spearman-Brown formula, reliabilities of .79, .84, and .88 for tests of length $1.12t$, $1.54t$, and $2.11t$, respectively. These values are slightly higher than those empirically obtained in this study for $T = 8$, 16, and 32 respectively. Thus, it can be argued that lengthening the test by simply adding additional MC items would have had a more beneficial impact on its reliability. However, this argument must be qualified. Such a process involves additional costs associated with the writing and testing of these new items (note that for the most extreme case of $T = 2.11t$, test length must be more than doubled). It is doubtful that this additional cost could be justified given the limited gain expected in terms of reliability. The notion that, given a fixed testing time, the test can be altered by manipulating either the number of distractors or of items is known as the assumption of proportionality. Recent empirical evidence (Budescu & Nevo, 1985) does not support this assumption, and indicates that reliability is better served by increasing the number of distractors.

Consider now the validity of the experimental tests with respect to the total score. Test theory predicts validities of .65, .67, and .68 for the three lengthened tests (e.g., Horst, 1966, p. 310). The values achieved for $T = 16$ and $T = 32$ equal or surpass these theoretical predictions; thus the present procedure appears to meet the expectations in terms of validity.

Wainer (1983) suggested that MM tests may help alleviate the guessing problem. What evidence is there that this effect is in fact achieved? Obviously, it is impossible to classify each response with certainty as either a "true" correct answer or a guess. However, several aspects of the data indicate that guessing is in fact reduced.

First, the data show a systematic decrease of the mean scores as a function of the total number of distractors. This result is consistent with the notion that the error component of the scores was reduced. Also, it was shown that the relationships between the test components (as measured by the test's measures of internal consistency, communality, and unidimensionality) improve as the number of distractors increases. The validity with respect to other verbal tests increases, but the correlations with tests of different factorial composition (quantitative) decrease. These results can be interpreted as indicators that MM tests eliminate part of the error variance inherent in the standard MC format. The item analysis provides direct evidence that the incorrect responses are distributed over a larger number of distractors, and that the probability of selecting each of the "new" distractors is close to its chance expectation. Finally, the examinees' responses to the feedback questionnaire indicate that they find it more difficult to guess in the presence of multiple distractors.

All of these results indicate that the MM tests appear to achieve their goal and that their construction is feasible. It appears that in order to benefit most from this format, the total number of distractors must be increased considerably. The best results in this study were achieved when $T = 16$ or 32. The discovery of the optimal composition of MM tests must be addressed in future research.

Finally, it is important to qualify this conclusion by pointing out that vocabulary tests, such as the one used here, are almost ideal for the MM format. These tests are almost pure power tests, with large numbers of short items and short answers. These characteristics tend to encourage guessing, and MM appears to be a viable solution to this problem. On the other hand, in tests with fewer and longer items and options, MM may not work as efficiently. In these cases guessing may not be such a salient problem, and the length of time required to scan and evaluate all of the options may be too heavy a price to pay in order to eliminate the guessing component.

Appendix

Mean and Variance of the Number of Chance Matches Under the MM Procedure

Assume a test of length n with a options per item. Further assume that a given examinee is certain of exactly K answers ($0 \leq K \leq n$). To complete the test, and assuming no omissions, he/she randomly selects ($n - K$) options from the remaining ($na - K$), without replacement. Because all selection sequences are equally probable, each option has a $1/(na - K)$ probability of being selected as a correct answer for each item. Define a random variable:

$$X_i = \begin{cases} 1 \text{ if match for item } i \\ 0 \text{ otherwise,} \end{cases}$$

such that the probability of a match is

$$P_r(X_i = 1) = \frac{1}{na-K} \qquad (i = 1, \ldots, n-K) \quad . \tag{A1}$$

Then define S_3, the number of matches in the sequence:

$$S_3 = \sum_{i=1}^{n-K} X_i \quad . \tag{A2}$$

Its expectation is simply

$$\mathrm{E}(S_3) = \sum_{i=1}^{n-K} \frac{1}{na-K} = \frac{n-K}{na-K} \quad . \tag{A3}$$

To derive the variance of S_3 note that

$$\mathrm{Var}(X_i) = \frac{1}{na-K} - \frac{1}{(na-K)^2} = \frac{na-K-1}{(na-K)^2} \quad , \tag{A4}$$

and also

$$\mathrm{Cov}(X_i,X_j) = \frac{1}{(na-K)(na-K-1)} - \frac{1}{(na-K)^2} = \frac{1}{(na-K)^2(na-K-1)} \quad . \tag{A5}$$

Finally, the variance of the number of correct chance matches is

$$\begin{aligned} \mathrm{Var}(S_3) &= (n-K)\mathrm{Var}(X_i) + (n-K)(n-K-1)\mathrm{Cov}(X_i,X_j) \\ &= \frac{(n-K)[(na-K-1)^2 + (n-K-1)]}{(na-K)^2(na-K-1)} \quad . \end{aligned} \tag{A6}$$

When $a = 1$, $\mathrm{E}(S_3) = \mathrm{Var}(S_3) = 1$, giving the results reported by Feller (1968) and Zimmerman and Williams (1982) for the special case of the SM task.

References

Abu-Sayf, F. K. (1977). A new formula score. *Educational and Psychological Measurement, 37,* 853–862.

Baldauf, R. B., Jr. (1982). The effects of guessing and item dependence on the reliability and validity of recognition based cloze tests. *Educational and Psychological Measurement, 42,* 855–867.

Baldauf, R. B., Jr., & Propst, I. K., Jr. (1979). Matching and multiple-choice cloze tests. *Journal of Educational Research, 72,* 321–326.

Budescu, D. V., & Nevo, B. (1985). Optimal number of options: An investigation of the assumption of proportionality. *Journal of Educational Measurement, 22,* 183–196.

Carlson, S. B. (1985). *Creative classroom testing.* Princeton NJ: Educational Testing Service.

Cross, H. L., & Frary, R. B. (1977). An empirical test of Lord's theoretical results regarding formula scoring of multiple-choice tests. *Journal of Educational Measurement, 14,* 313–321.

Diamond, J., & Evans, W. (1973). The correction for guessing. *Journal of Educational Research, 43,* 181–191.

Feller, W. (1968). *An introduction to probability theory and its applications, Vol. 1* (3rd ed.). New York: Wiley.

Frary, R. B., & Hutchinson, T. P. (1982). Willingness to answer multiple-choice questions as manifested both in genuine and in nonsense items. *Educational and Psychological Measurement, 42,* 815–821.

Gulliksen, H. (1986). Perspective on educational measurement. *Applied Psychological Measurement, 10,* 109–132.

Horst, P. (1966). *Psychological measurement and prediction.* Belmont CA: Wadsworth.

Hutchinson, T. P. (1982). Some theories of performance in multiple choice tests, and their implications for variants of the task. *British Journal of Mathematical and Statistical Psychology, 35,* 71–89.

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

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale NJ: Erlbaum.

Molenaar, W. (1977). On Bayesian formula scores for random guessing in multiple-choice tests. *British Journal of Mathematical and Statistical Psychology, 30,* 79–89.

Reilly, R. R. (1975). Empirical option weighting with a correction for guessing. *Educational and Psychological Measurement, 35,* 613–619.

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

Slakter, M. J. (1969). Generality of risk taking on objective examinations. *Educational and Psychological Measurement, 29,* 115–128.

Smith, J. K. (1982). Converging on correct answers: A peculiarity of multiple choice items. *Journal of Educational Measurement, 19,* 211–220.

Swineford, F. (1941). Analysis of a personality trait. *Journal of Educational Psychology, 32,* 438–444.

Traub, R. E., Hambleton, R. K., & Singh, B. (1969). Effects of promised reward and threatened penalty on performance of a multiple-choice vocabulary test. *Educational and Psychological Measurement, 29,* 847–861.

van der Ven, A. H. G. S. (1974). A Bayesian formula score for the simple knowledge or random guessing model. *Ned. Tijdchr. Psychol., 29,* 409–414.

Wainer, H. (1983). Are we correcting for guessing in the wrong direction? In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 63–80). New York: Academic Press.

Wesman, A. G. (1971). Writing the test item. In R. L. Thorndike (Ed.), *Educational measurement.* Washington DC: American Council on Education.

Wood, R. (1984). Review of "New horizons in testing: Latent trait test theory and computerized adaptive testing". *Applied Psychological Measurement, 8,* 463–465.

Ziller, R. C. (1957). A measure of the gambling response-set in objective tests. *Psychometrika, 22,* 289–292.

Zimmerman, D. W., & Williams, R. H. (1982). Element of chance and comparative reliability of matching tests and multiple-choice tests. *Psychological Reports, 50,* 975–980.

Zinger, A. (1972). A note on multiple-choice items. *Journal of the American Statistical Association, 67,* 340–341.

Acknowledgments

The author thanks B. Nevo and Y. Cohen of the National Institute of Testing and Evaluation for assistance in the collection of the data, and Danny Shteinberg for his help in the preparation of the tests and the data analysis.

Author's Address

Send requests for reprints or further information to David Budescu, Department of Psychology, University of Haifa, Haifa 31999, Israel.