Ordering Power of Separate versus Grouped True-False Tests: Interaction of Type of Test with Knowledge Levels of Examinees

Louis M. Hsu
Fairleigh Dickinson University

The ordering power of an objective test was defined in terms of the probability that this test led to the correct ranking of examinees. A comparison of the relative ordering power of separate and grouped-items true-false (T-F) tests indicated that neither type of test was uniformly superior to the other across all levels of knowledge of examinees. Instead, separate-items T-F tests were found to be superior in discriminating among examinees with medium and high levels of knowledge, and grouped-items T-F tests with two and three items per cluster were found to be superior for discriminating among examinees with low levels of knowledge. These findings do not support blanket recommendations such as Ebel's (1978) that "test constructors should avoid constructing items in multiple-choice form which are essentially collections of T-F statements" (p. 43) or that, in general, "it is better to present such statements as independent T-F items" (p. 43). Rather, they are similar to Lord's (1977) findings concerning the relative efficiency of multiple-choice tests with different numbers of options per question for examinees of differing ability levels.

This paper examines the relative merits of ordinary separate-items true-false tests (S tests) and grouped-items true-false tests (G tests) as a function of knowledge levels of examinees. An S test is composed of a set of questions, each of which consists of one true-false (T-F) item. The examinee's task is to determine the truth value (T or F) of each item. A G test consists of a set of questions, each of which comprises a group or cluster of T-F items. Each cluster is introduced by a stem that asks essentially, "Which of the following is true [or false]?" (Ebel, 1978, p. 37). The only requirement for such clusters is that only one of the statements should be true (or false).

Ebel (1978) has recently noted some a priori grounds for expecting S tests to be superior to G tests: Each T-F item presented separately yields an independent indication of ability. When grouped, however, the several T-F items in a group yield only one indication of ability. Separate presentation of T-F items provides more information. Ebel (1978) noted that test specialists have suggested that G items are undesirable.

Ebel (1978) also noted some a priori grounds for expecting G tests to be superior to S tests: The principal argument used in favor of grouping is that S-test items are more susceptible to errors from

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 3, No. 4 Fall 1979 pp. 529–536
© Copyright 1979 West Publishing Co.

guessing. For example, grouping statements in threes would reduce from .5 to .33 the probability of a correct response from blind guessing.

An empirical study (Ebel, 1978) involving S and G tests made up of the same T-F items led Ebel to confirm the expectation on rational grounds that G-test items would be less effective than S-test items in measuring achievement. The principal result leading Ebel to this conclusion appears to be that the reliability coefficients for the S tests were substantially higher than for the G tests. He noted that two inferences could be drawn from the higher reliability of the S tests: (1) the separate items did in fact provide a greater quantity of dependable information about differences in examinee achievements and (2) S-test items were not subject to substantial guessing error. Ebel's position thus appears to be that the greater *overall reliability* of S over G tests indicates the superiority of S tests over G tests.

Recently, in a paper on a related topic Lord (1977) noted that "... it is not enough to just think about overall test reliability ..." (p. 36). This comment was based on his theoretical finding that the effect of decreasing the number of choices per item, while lengthening the test proportionately, is to increase the efficiency (defined in terms of the length of the asymptotic confidence interval for estimating ability from test scores; Birnbaum, 1968) of the test for high-level examinees and to decrease its efficiency for low-level examinees. Lord (1977) explained these findings as follows: At high ability levels, there is little random guessing, and the relative efficiency of multiple-choice tests differing in terms of the number of options included per question is determined by their length; and at low ability levels, the effect of random guessing becomes of overwhelming importance. Lord (1977) summarized his findings by noting that "... for any pair of tests [differing in terms of number of options per question], one test is better than the other over a certain range of ability, but worse over the complementary range of ability" (p. 36).

Due to the similarity of multiple versus separate T-F tests on the one hand and many-option-versus few-option-multiple-choice tests on the other, it would not be surprising if results similar to those of Lord were obtained in comparing the characteristics of S and G tests for examinees differing in knowledgeability. The following intuitive argument would lead to an expectation of interaction of test type (S versus G) with knowledge levels of examinees: The stability of test scores of examinees (over repeated testings with parallel forms of tests) would be expected to decrease with decrease in knowledge levels of examinees for both S and G tests. This would be expected because of the effect of increased guessing with decrease in knowledge levels. However, in view of the difference in the number of T-F items per question of S and G tests, the *rate* of decrease in stability of scores of S and G tests, with decrease in knowledge levels of examinees, would be expected to be different: hence, the anticipated interaction. The model and derivations of the following sections develop this argument more precisely.

The criterion to be used to evaluate S and G tests at different knowledge levels of examinees will be called "ordering power." The ordering power of a test is defined in terms of the extent to which the test allows correct ranking of examinees with respect to the measured characteristic. More specifically, the ordering power of a test for a pair of examinees will refer to the probability that the test leads to a correct ranking of these examinees with respect to the measured property.

The object of the present paper is to demonstrate that the relative ordering power of S and G tests depends on the ability (or knowledge) levels of the examinees. More specifically, S tests discriminate better than G tests at some levels of ability, and G tests discriminate better than S tests at other levels of ability. An assumption underlying the findings of the present paper is that the total number of T-F items included in the S and G tests to be compared is constant. This assumption was made by Ebel (1978) in his discussion of the ineffectiveness of G tests, and a similar assumption was made by Grier

(1975) and Lord (1977) in their discussion of the optimal number of choices per item in multiple-choice tests. As noted by Lord (1977), this assumption makes sense if the total testing time for a set of r items is proportional to the number of choices per item. Methods of comparing the ordering power of S and G tests in situations in which this assumption is not tenable will also be discussed.

The Model

The model postulated in this paper involves sampling from a T-F item universe. The total number of T-F items in both the S and G tests will be designated "nr." The number of items per cluster will be designated "n," and the number of clusters in the G test will be designated "r." Both the nr T-F items included in the S tests and the nr T-F items included in the G test are viewed as random samples from a very large universe (relative to the size of the T-F tests) of T-F items. π_A designates the proportion of items in this universe whose truth values are known by Examinee A; π_B designates the proportion of items in this universe whose truth values are known by Examinee B. π will, in general, be referred to as the level of knowledge, or the ability level, of the examinee (see Wilcox, 1976).

The classical assumption (Guilford, 1954, p. 484) that an examinee, when presented with an item whose truth value he/she does not know, guesses its truth value randomly will be considered tenable. Although this assumption has been criticized as unrealistic in many situations (Ebel, 1968; Lord & Novick, 1968, p. 309), it has more recently been judged "not likely to lead to unreasonable conclusions" (Lord, 1977, p. 34) when dealing with a problem analogous to the one discussed in the present paper, viz., the problem of determining the optimal number of options to be used per question in multiple-choice tests. The classical assumption has also been described as a desirable property of objective tests (Weitzman, 1970), and methods have been proposed for determining when items in objective tests have this property (Weitzman, 1970).

Given the above model, it is clear that the probability that an examinee will correctly answer a T-F item drawn at random from the universe would be

$$P(\text{item known})\,P(\text{correct} \mid \text{item known}) + P(\text{item unknown})\,P(\text{correct} \mid \text{item unknown}) =$$

$$\{\ (\pi)(1) + (1-\pi)(\tfrac{1}{2})\ \}. \qquad [1]$$

Similarly, when T-F items for a cluster of n items consisting of $(n-1)$ items of one type (say T) and 1 item of the other type (say F) are drawn randomly from respective item universes of T and F items, in which the examinee knows answers to π of the T items and also π of the F items, the probability that the examinee will choose the correct item in the cluster would be

$$\begin{aligned}&P(n \text{ items known})\,P(\text{correct} \mid n \text{ items known}) +\\ &P(n-1 \text{ items known})\,P(\text{correct} \mid n-1 \text{ items known}) +\\ &P(n-2 \text{ items known})\,P(\text{correct} \mid n-2 \text{ items known}) +\\ &\quad\vdots\\ &P(0 \text{ items known})\,P(\text{correct} \mid 0 \text{ items known}) =\end{aligned}$$

$$\pi^{n}(1) + {}_{n}C_{n-1}\,\pi^{n-1}(1-\pi)(1) + {}_{n}C_{n-2}\,\pi^{n-2}(1-\pi)^{2}((n-1)/n)$$

$$+ (1-\pi)^{n}(1/n). \qquad [2]$$

More concisely, the probability that the examinee will choose the correct item in the cluster would be

$$\pi^{n} + \sum_{i=1}^{i=n} {}_{n}C_{(n-i)} \, \pi^{(n-i)} \, (1-\pi)^{i} \, \{(n-i+1) \,/\, (n)\}, \qquad [3]$$

where $_{n}C_{n-i} = n!/[(n-i)!\ (i)!]$. The general term of the expression on the right is obtained by calculating

$P\{(n-i$ known items) $\cap$ (answer is one of these) $\cap$ (correct answer)$\}$ +
$P\{(n-i$ known items) $\cap$ (answer is not one of these) $\cap$ (correct answer)$\}$

which may be rewritten, factoring out $P(n-i$ known items), as

$${}_{n}C_{(n-i)} \, \pi^{(n-i)}(1-\pi)^{i} \, \{[(n-i)/(n)][1] + [i/n][1/i]\} \qquad [4]$$

or, equivalently, as

$${}_{n}C_{(n-i)} \, \pi^{(n-i)} \, (1-\pi)^{i} \, \{(n-i+1) \,/\, (n)\}. \qquad [5]$$

Thus, the expected proportions of correct responses by an examinee with level of knowledge π is $E(p_S)$ = Equation 1 for the S test and $E(p_G)$ = Equation 3 for the G test consisting of r clusters, each containing n T-F items (where p_S = proportion of correct responses with the S test, p_G = proportion of correct responses with the G test, and i = number of items in a cluster whose answers are unknown by the examinee).

Considering that an examinee who chooses the correct T-F item in a cluster will be given credit for knowing all items in the cluster (and that otherwise he/she will be given credit for none), and considering that (nr) items are to be included in both the S and G test, it is clear that the variances of the number of correctly answered T-F items (X) by the examinee would be

$$\begin{aligned} V(X_S) &= nr\,[\{\tfrac{1}{2}(1+\pi)\}\,\{1-\tfrac{1}{2}(1+\pi)\}^2 + \{1-\tfrac{1}{2}(1+\pi)\}\,\{0-\tfrac{1}{2}(1+\pi)\}^2] \\ &= nr\,\{E(p_S)\}\,\{1 - E(p_S)\} \end{aligned} \qquad [6]$$

for the S test and

$$\begin{aligned} V(X_G) &= r\,[\,\{E(p_G)\}\{n - E(p_G)\}^2 + \{1 - E(p_G)\}\{0 - E(p_G)\}^2\,] \\ &= r\,[E(p_G)\,\{1 - E(p_G)\}] \end{aligned} \qquad [7]$$

for the G test.

Examinees A and B will be correctly ranked using a test if the rank order of their scores on that test is the same as the rank order of their levels of knowledge. Since their scores (X) can be expected to be approximately normal random variables, provided that π levels are not too large and r is not too small, the probability that Examinees A and B, whose performances on the test are assumed to be independent, will be correctly ranked (given that $\pi_A > \pi_B$) can therefore be approximately determined by reference to the Z-normal distribution, i.e.,

$$P\left[z > z_S = \frac{(nr\ E(p_S)_A - nr\ E(p_S)_B)}{\sqrt{V(X_S)_A + V(X_S)_B}} \right], \qquad [8]$$

which may also be written as

$$P\left[z > z_S = \sqrt{r}\ \frac{(n\ E(p_S)_A - n\ E(p_S)_B)}{\sqrt{n\ E(p_S)_A(1-E(p_S)_A) + n\ E(p_S)_B(1-E(p_S)_B)}} \right] \qquad [9]$$

for an S test and approximately

$$P\left[z > z_G = \frac{(r\ E(p_G)_A - r\ E(p_G)_B)}{\sqrt{V(X_G)_A + V(X_G)_B}} \right], \qquad [10]$$

which may also be written as

$$P\left[z > z_G = \sqrt{r}\ \frac{(E(p_G)_A - E(p_G)_B)}{\sqrt{E(p_G)_A(1-E(p_G)_A) + E(p_G)_B(1-E(p_G)_B)}} \right] \qquad [11]$$

for a G test (where z is approximately normally distributed).

Determination of the Ordering Power of S and G Tests

Equations 8 and 10 were applied to S and G tests with $(n \times r) = 120$ T-F items, assuming that $\pi_A - \pi_B = .1$, with mid-values ranging from .15 to .85, in steps of .1. Table 1 lists obtained z_S and z_G values, and Table 2 lists corresponding probabilities of correctly ranking Examinees A and B. It is clear from examination of Tables 1 and 2 that the S test with 120 items was superior to the G test with 2, 3, 4, and 5 items per cluster (and a total of $(r \times n) = 120$ T-F items) in discriminating small (about .1) differences in levels of knowledge when examinees' levels of knowledge are medium or high (say, above approximately .35). It is also clear, however, that G tests with 3 and 2 items per cluster had greater ordering power than S tests for low levels of ability (say, below about .35). Thus, from Tables 1 and 2, it is apparent that with respect to ordering power, S tests with 120 items were not uniformly superior to G tests with nr=120, across ability levels of examinees.

Generalization to Cases in Which the Number of T-F Items is (krn) for both S and G Tests

It is clear, then, that the relative ordering power of the S and G tests for a pair of examinees is determined by the relative sizes of z_S and z_G in Equations 8 and 10. That is, if and only if the ratio

Table 1
z_S and z_G Values for S and G Tests with 120 T-F Items

Number of T-F Items in Cluster	z from Eqs. 8 or 10	Mid-Value of π							
		.15	.25	.35	.45	.55	.65	.75	.85
Separate									
1	z_S	.78	.80	.83	.87	.93	1.02	1.17	1.48
Grouped									
2	z_G	.97	.92	.87	.84	.81	.79	.77	.74
3	z_G	.88	.85	.84	.83	.83	.83	.83	.82
4	z_G	.79	.76	.76	.76	.78	.80	.82	.83
5	z_G	.73	.70	.69	.70	.72	.75	.79	.82

z_G/z_S is smaller than 1 will the S test yield a larger probability of a correct ranking than the G test for any fixed pair of π values.

An interesting fact which may be observed from examination of Equations 8 and 10 is that the value of the ratio (z_G/z_S) for a given pair of examinees is independent of the number of clusters (r) included in the G Test (given that the total number of T-F items, rn, is constant). This is apparent in Equations 9 and 11, which are the equivalents of Equations 8 and 10, respectively. Thus, it is possible to determine which test (S or G) is more likely to lead to a correct ranking of two examinees for any fixed rn (such as rn = 120, above) and to generalize the conclusions to any contrast of S and G tests which contain $k(rn)$ T-F items and the same number of T-F items per cluster. The only requirement is that k be sufficiently large for the normal approximation method to be applicable.

Considering that two examinees (A and B) differ by .1 in the proportion of items whose truth value they know in the item universe (i.e., $\pi_A - \pi_B = .1$), values of the ratio (z_G/z_S) were calculated for average levels of knowledge ranging from .15 to .85 for S tests and for G tests with 2, 3, 4, and 5 items per cluster. These values appear in Table 3. Note that the ratios were obtained from the z_S and z_G

Table 2
Probabilities of Correct Rankings of Examinees with 120 Item S and G Tests

Number of T-F Items in Cluster	Mid-Value of π							
	.15	.25	.35	.45	.55	.65	.75	.85
1	.782	.788	.797	.808	.824	.846	.879	.931
2	.834	.821	.808	.800	.791	.785	.779	.770
3	.811	.802	.800	.797	.797	.797	.797	.794
4	.785	.776	.776	.776	.782	.788	.794	.797
5	.767	.758	.755	.758	.764	.773	.785	.794

values in Table 1. Clearly, however, the same ratios would have been obtained if $(n \times r) = 240$ or, more generally, if $(n \times r) = k(120)$. It is apparent from examination of Table 3 that S tests can be expected to have greater ordering power (i.e., to lead to more correct rankings) than G tests for examinees with medium and high levels of knowledge (say, π of at least .35) but that G tests with 2 or 3 items per cluster can be expected to have greater ordering power than S tests for examinees with low levels of knowledge (say, π less than 0.35), given that the total number of T-F items (i.e., $n \times r$) is equal in the S and G tests. The data in Tables 1 and 3 were generated assuming that the difference in levels of knowledge of the two examinees was .1. Clearly, however, similar results would be obtained with somewhat larger or smaller values of $\pi_A - \pi_B$.

Generalization to Cases in Which the Number of T-F Items in the S Test is not the Same as that in the G Tests

The data in Table 3 and the conclusions drawn from these data were generated assuming that the total number of T-F items $(n \times r)$ was the same for the S and G tests. This assumption would be justified if the total testing time was proportional to the number of choices per question (Lord, 1977). In situations in which this is not the case, it is clear that the method of assessing the relative ordering power of S and G tests developed in this paper could easily be appropriately modified. The modification would involve changing the ratio of the number of T-F items in the S test $\{(nr)_S\}$ to the number of T-F items in the G test $\{(nr)_G\}$ in Equations 8 and 10 so as to reflect the ratio of the number of separate T-F items completed in a fixed period of time to the number of clusters of items completed in the period by the examinees.

For example, Frisbie (1973) collected empirical data on this ratio and found that students answered 1.5 times as many T-F items as four-alternative-multiple-choice items (dealing with social and natural science) in a fixed period of time. Thus, in assessing the relative ordering power of S tests and G tests with four items per cluster (dealing with this content area), the ratio $\{(nr)_S/(nr)_G\}$ would be kept at about 1.5 in calculating z_S and z_G values with Equations 8 and 10.

It is apparent that the most realistic ratio $\{(nr)_S/(nr)_G\}$ in any given situation involving the comparison of the discriminative capacity of S and G tests would be contingent on the content area

Table 3
z_G/z_S Values Indicating the Relative Ordering Power of Grouped vs. Separate True-False Tests for Examinees with Different Levels of Knowledge

Number of T-F Items in Cluster	Mid-Value of π							
	.15	.25	.35	.45	.55	.65	.75	.85
1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2	1.24	1.15	1.05	0.97	0.87	0.77	0.66	0.50
3	1.13	1.06	1.01	0.95	0.89	0.81	0.71	0.55
4	1.01	0.95	0.92	0.87	0.84	0.78	0.70	0.56
5	0.94	0.88	0.83	0.80	0.77	0.74	0.68	0.55

sampled (Frisbie, 1973) and on the number of T-F items to be included per cluster in the G test. Once this ratio has been determined (as was done by Frisbie, 1973), Equations 8 and 10 could be applied to determine the ranges of knowlege of examinees for which S and G tests, respectively, are preferable.

Conclusions

The findings reported in the present paper do not support blanket recommendations such as Ebel's (1978) that ". . . test constructors should avoid constructing items in multiple-choice form which are essentially collections of T-F statements" (p. 43) or that, in general, "it is better to present such statements as independent T-F items" (p. 43). Instead, the present findings suggest that when S and G tests contain the same number of T-F items, G tests are preferable to S tests in situations in which the examinees have low levels of knowledge or in situations in which discrimination among such individuals is of more importance than among other individuals. When the number of T-F items in S and G tests are not equal, it is clear that by applying Equations 8 and 10, using appropriate empirically determined ratios of $(nr)_S/(nr)_G$, comparable conclusions could be drawn.

That is, in general, it would be expected that S tests would be better than G tests for discriminating among examinees in one range of levels of knowledge and that G tests would be preferable to S tests in discriminating among examinees in a complementary range of levels of knowledge. These conclusions are analogous to Lord's (1977) expectations concerning the optimal number of choices per question in multiple-choice tests: That is, just as the optimal number of choices per question in multiple-choice tests is dependent on the ability levels of the examinees, so also the optimal number of T-F items per question (viz., one or some number greater than one) is dependent on the knowledge levels of the examinees.

References

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley, 1968.

Ebel, R. L. Blind guessing on objective achievement tests. *Journal of Educational Measurement,* 1968, *5,* 321-325.

Ebel, R. L. The ineffectiveness of multiple true-false test items. *Educational and Psychological Measurement,* 1978, *38,* 37-44.

Frisbie, D. A. Multiple-choice versus true-false: A comparison of reliabilities and concurrent validities. *Journal of Educational Measurement,* 1973, *10,* 297-304.

Grier, J. B. The number of alternatives for optimum test reliability. *Journal of Educational Measurement,* 1975, *12,* 109-113.

Guilford, J. P. *Psychometric methods.* New York: McGraw-Hill, 1954.

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

Lord, F. M., & Novick, M. R. *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley, 1968.

Weitzman, R. A. Ideal multiple-choice items. *Journal of the American Statistical Association,* 1970, *65,* 71-89.

Wilcox, R. A note on the length and passing score of a mastery test. *Journal of Educational Statistics,* 1976, *1,* 359-364.

Acknowledgments

The author thanks the Editor and two anonymous reviewers for their helpful comments on an earlier draft of this paper.

Author's Address

Louis M. Hsu, Department of Psychology, Fairleigh Dickinson University, Teaneck, NJ 07666.