A Meta-Analytic Investigation of the Effect of Various Test Item Characteristics on Test Scores and Test Completion Times

The current study reported the results of a meta-analytic investigation of the effects on test scores and test completion times of three aspects of writing test items: The number of answers in multiple-choice exams, the order of item difficulty, and the organization of items by content. The results of meta-analysis indicated that three-choice questions are slightly easier than four-choice questions (d = .90) and take significantly less time to complete (d = -.61). Exams beginning with easier items and then moving to more difficult items are slightly easier than exams with randomly ordered items (d = .11) or exams beginning with difficult items (d = .22). Exams in which the items are organized by content are slightly easier than exams containing randomly ordered items (d = .04). All of the above effect sizes are small.

By
Michael G. Aamodt
Teige McShane

Creating selection and promotion exams are common tasks for individuals engaged in public personnel management. Tests that are properly created are content valid, psychometrically sound, and discriminate between applicant levels of knowledge.

Though the goals of good tests are clear, creating the ideal test is not easy. One of the biggest problems in creating exams involves including all of the items needed to cover a job related area, but at the same time, keeping the test at a reasonable length.

One solution to this problem is to reduce the length of a multiple-choice exam by using only three choices for each item rather than the traditional four or five. For example, the traditional four choice question would change from:

What is the average validity of ability tests?
a) .51 b) .20 c) .00 d) .40

to:

What is the average validity of ability tests?
a) .51 b) .20 c) .00

Michael G. Aamodt received his Ph.D. from the University of Arkansas and is currently an associate professor of industrial/organizational psychology at Radford University in Virginia.

The idea here is that by using only three choices, the items will not only be easier to write, but will shorten the time taken to complete the exam by reducing the amount of material that a test taker needs to read and analyze.

Critics to such a solution believe that by reducing the number of distractors, exam scores will increase because a test taker who does not know the answer to a question will have a one-in-three chance of correctly guessing the answer compared to the one-in-four chance that would be associated with four choices.

Another problem in creating exams involves the need for multiple forms. For example, in group testing situations, multiple forms of the same test are needed to help reduce the opportunities for cheating. Typically, multiple forms are created in one of two ways.

The first method is to write completely new items covering the same content area. The problem with this method is that good items are difficult and time consuming to write. Consequently, the alternative strategy is to use the same test times, but to rearrange them in a different order. Critics of this last strategy claim that changing the order of items may affect exam scores. That is, an exam with a larger proportion of easy items early in the exam will be easier than an exam with a larger proportion of difficult items early in the exam. Or, as shown in Figure 1, an exam in which the items are arranged by content might be easier than an exam in which the items are randomly ordered because the test taker will be able to maintain one train of thought and can take advantage of self-addressable memory (recalling information the same way in which it was stored in memory).

The purpose of this paper is to report the results of a meta-analysis of the research that has been conducted on the effects of both the number of choices to multiple choice questions and the arrangement of test items in an exam. A meta-analysis is a statistical method of cumulating varied research results into one overall effect size.

Method

Finding Research Articles

The first step in a meta-analysis is to find research articles on the topic of interest. Articles for this meta-analysis were located by using two written sources; the Social Sciences Index and the Business Periodical Index, and two computerized sources; Infotrac and Silver Platter. The reference lists from the articles identified from the above sources were also used to locate additional articles.

Teige McShane received his M.A. from Radford University and is currently a Human Resoources Representative for Purdue Farms, Inc. in North Carolina.

To be included in this meta-analysis the article had to report the results of an experiment and had to include a statistic that could be converted into an effect size. This procedure resulted in eight studies with 14 samples investigating number of choices, 20 studies with 26 samples investigating item order based on item difficulty, and four studies with 16 samples investigating the organization of items based on item content. These articles are referenced in Appendix A.

Converting Research Findings to Effect Sizes

Once the research articles had been located, the statistical result for each study was converted into an effect size using the formulas provided in Mullin and Rosenthal (1985) and with the help of the computer program developed by Johnson (1989). A few studies did not provide statistical information that could be converted into an effect size and thus were not included in the analysis.

Cumulating Effect Sizes

After the individual effect sizes had been computed, the effect size for each study was weighted by the size of the sample and combined using the method suggested by Hunter, Schmidt, and Jackson (1982) and the computer program developed by Schwarzer (1989). In addition to a mean effect size ($\overline{d}$), the observed effect size variance, the amount of variance expected due to sampling error, and a 95% confidence interval were calculated.

Results and Discussion

Number of Choices

As shown in Table 1, the effect for the number of choices on exam scores was only .09 with a confidence interval ranging from .04 to .14. An effect size can be interpreted in terms of standard deviations. For example, an effect size of 1.0 would indicate that a particular variable will change the dependent variable by a full standard deviation. In this case, the presence of three choices rather than four would increase test scores by .09 of a standard deviation. In the articles reviewed, the mean test score on a 100-item exam was 67.8 with a standard deviation of 13.5. Multiplying this standard deviation by the mean effect size of .09 indicates that exam scores would only increase by 1.22 points if three choices were used rather than four or five.

As also shown in Table 1, the mean effect on the time taken to complete an exam was reduced by using three choices. The mean time taken to complete a 100 item exam across the four studies in our sample

Table 1
Meta-Analysis Results for Number of Choices

	Dependent Variable		
	Exam Score	Completion Time	Item Discrimination
Number of Studies	14	4	11
Sample Size	6,789	674	4,263
Mean Effect Size	.09	-.61	.05
Observed Variance	.04	.30	.05
Sampling Error	.02	.03	.02
95% Lower Bound	.04		-.45
95% Upper Bound	.13		-.77

was 37.2 minutes with a standard deviation of 7.52. Multiplying the mean effect size of -.61 by the standard deviation of 7.52 reveals a time savings of 4.59 minutes. Because test takers were able to complete 2.69 items per minute, this savings suggests that with the use of three rather than four choices, 12.4 more items could be added to an exam and completed in the same amount of time as the 100 item four-choice exam. This increase in items without in increase in time may result in a more content valid exam.

As also shown in Table 1, using three choices rather than four did not reduce the ability of an exam to discriminate between low and high scorers. In fact, if anything, item discrimination was better with three rather than four choices.

In meta-analysis, any time a confidence interval includes zero, one is not sure that the variable in question actually had an effect on the dependent variable because the mean effect size could actually be zero. In this case, the confidence interval does include zero indicating that the number of choices may not have an effect on item discrimination.

Item Difficulty Order

As shown in Table 2, there was a small difference between test scores of exams with the easier items located early in the exam and the more difficult items located later in the exam versus exams in which the item order was random. Even had the confidence interval not included zero, the effect for item order was small as the easy-to-hard order would only improve test scores by 1.49 points on a 100-item exam.

As shown in Table 3, the easy-to hard order did not significantly lower the anxiety level of test takers compared to a random order arrangement. As shown in Table 4, the easy-to-hard order also did not significantly affect the perceived difficulty of the exam.

Table 2
Meta-Analysis Results for Effect of Item Order on Test Scores

	Item Order Comparison		
	Easy to Hard versus Random	Easy to Hard versus Hard to Easy	Hard to Easy versus Random
Number of Studies	26	11	3
Sample Size	13,438	1,564	297
Mean Effect Size	11	.22	-.08
Observed Varianc	.09	.05	.02
Sampling Error	.04	.03	.04
95% Lower Bound	.00	.12	-.08
95% Upper Bound	.22	.32	-.08

Table 3
Meta-Analysis Results for Effect of Item Order on Anxiety Level

	Item Order Comparison	
	Easy to Hard versus Random	Easy to Hard versus Hard to Easy
Number of Studies	5	3
Sample Size	378	189
Mean Effect Size -	.08	-.30
Observed Variance	.16	.01
Sampling Error	.08	.01
95% Lower Bound	-.44	-.30
95% Upper Bound	.23	-.30

As shown in the second column of Table 2, exams beginning with easy items were significantly easier than exams beginning with difficult items. The effect size of .22 translates to a difference of about 2.97 points on a typical 100 item exam. As shown in Table 3, the easy-to-hard order was also less anxiety producing than the hard-to-easy order.

Beginning an exam with hard items rather than randomly arranged items results in a small but significant reduction in exam scores. Because the amount of variance expected by sampling error alone was greater than the amount of observed variance, the upper and lower bounds of the 95% confidence interval are the same as the mean effect size. The effect size of -.08 corresponds to a difference of about 1.08 points on a typical 100 item exam.

Table 4
Meta-Analysis Results for Effect of Item Order on Perceived Difficulty

	Item Order Easy to Hard versus Random
Number of Studies	9
Sample Size	848
Mean Effect Size -	.33
Observed Variance	.36
Sampling Error	.07
95% Lower Bound	-.72
95% Upper Bound	.06

Organization of Items by Content

As depicted in Table 5, there was a significant, but very small effect of item organization on exam scores. Based on the same exam figures used in previous examples, a test taker will score .54 of a point higher on an exam organized by content than he/she will on an exam which is randomly organized.

Table 5
Meta-Analysis Results for Effect of Item Organization on Test Scores

Number of Studies	16
Sample Size	3,731
Mean Effect Size	.04
Observed Variance	.01
Sampling Error	.01
95% Lower Bound	.04
95% Upper Bound	.04

Summary

As indicated in the results discussed above, neither the number of choices in an item nor the arrangement of items in an exam greatly affect exam scores. Thus, three-choice items can be confidently used with the advantage of less time to both create and take the exam.

The lack of large effects for item order indicate that the systematic rearrangement of the same items to form multiple forms of an exam (e.g., to prevent cheating in a group testing situation) will not greatly change the scores on an exam. It is important to keep in mind that the effect sizes for item order that were reported in this analysis are probably upper limits. For example, a person taking an exam with the easier items early in the exam will score about 1.5 point higher than a person taking the same exam with the items arranged in a random fashion. However, most multiple forms consist of randomly ordered items and the limited available research indicates that there are no differences between two tests with randomly ordered items.

If for some reason, it was desirable to make a test more difficult, beginning the test with the more difficult items, using four rather than three choices, and not organizing items by content would theoretically reduce the average score by about four or five points. Conversely, a test can also be made about three points easier by organizing the items by content, using three choices, and beginning the exams with the easier items.

Figure 1
Example of Content Organized Items and Randomly Organized Items

Content Organized items

Q1.	State Law
Q2.	State Law
Q3.	State Law
Q4.	Police Procedure
Q5.	Police Procedure
Q6.	Police Procedure
Q7.	Administrative Skills
Q8.	Administrative Skills
Q9.	Administrative Skills

Items with Random or Mixed Organizations

Q1.S	tate Law
Q2.	Police Procedure
Q3.	Administrative Skills
Q4.	State Law
Q5.	Administrative Skills
Q6.	Police Procedure
Q7.	Police Procedure
Q8.	State Law
Q9.	Administrative Skills

References

Hunter, J.E., Schmidt, F.L., & Jackson, G.B. (1982). *Meta-analysis and cumulating research findings across studies.* Beverly Hills, CA: Sage.

Johnson, B.T. (1989). *DSTAT software for meta-analytic review of research literatures.* Hillsdale, NJ: Lawrence Erlbaum.

Mullin, B., & Rosenthal, R. (1985). *Basic meta-analysis: Procedures and programs.* Hillsdale, NJ: Erlbaum.

Schwarzer, R. (1989). *Meta-analysis programs.* Durham, NC: National Collegiate Software Clearinghouse of Duke University Press. ·

Appendix A

Studies Used in the Meta-Analysis

Balch, W.R. (1989). Item order affects performance on multiple-choice exams. *Teaching of Psychology, 16* (2), 75-77.

Barcikowsky, R.S., & Olsen, H. (1975). Test item arrangement and adaptation level. *Journal of Psychology*, 90, 87-93.

Brenner, M.H. (1964). Test difficulty, reliability, and discrimination as functions of item difficulty order. *Journal of Applied Psychology*, 48, 98-100.

Burgess, T.C. (1968). Item arrangement and student evaluation of examinations. *Psychology, 5*, 69-73.

Capron, V.L. (1933). Relative effect of three orders of arrangement of items upon pupils scores in certain arithmetic and spelling tests. *Journal of Educational* Psychology, *24*, 687-694.

Costin, F. (1970). The optimal number of alternatives in multiple-choice achievement tests: Some empirical evidence for a mathematical proof. *Educational and Psychological Measurement, 30*, 353-358.

Costin, F. (1972). Three-choice versus four-choice items: Implications for reliability and validity of objective achievement tests. *Educational and Psychological Measurement, 32*, 1035-1038.

Costin, F. (1975). Difficulty and homogeneity of three-choice versus four-choice objective test items when matched for content of stem. *Teaching of Psychology*, 144-145.

Dambrot, F. (1980). Test item order and academic ability, or should you shuffle the test item deck? *Teaching of Psychology, 7*(2), 94-96.

Feder, D.D. (1936). Effect of direction and arrangements of items on students performance in a test. *Journal of Educational Research, 30*, 28-35.

Flaugher, R.L., Melton, R.S., & Myers, C.T. (1968). Item rearrangement under typical test conditions. *Educational and Psychological Measurement, 28*, 813-824.

Gerow, J.R. (1980). Performance on achievement tests as a function of the order of item difficulty. *Teaching of Psychology, 7*(2), 93-94.

Grier, J.B. (1975). The number of alternatives for optimum test reliability. *Journal of Educational Measurement, 12*(2), 109-113.

Hambleton, R.K., & Traub, R.E. (1974). The effects of item order on test performance and stress. *Journal of Experimental Education, 43*, 40-46.

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

Huck, S.W., & Bowers, N.D. (1972). Item difficulty level and sequence effects in multiple choice-achievement tests. *Journal of Educational Measurement, 9*, 105-111.

Klosner, N.C., & Gellman, E.K. (1973). The effect of item arrangement on classroom test performance: Implications for content validity. *Educational and Psychological Measurement, 33*, 413-418.

Laffitte, R.G. (1984). Effects of item order on achievement scores and students' perception of test difficulty. *Teaching of Psychology, 11*(4), 212-214.

Marso, R.N. (1970). Test item arrangement, testing time, and performance. *Journal of Educational Measurement, 7*, 113-118.

McShane, T.D., & McGown, M.C. 1989). Effect of the number of item distractors on test scores and completion times. *Proceedings of the 10th Annual Graduate Conference in I/O Psychology and Organizational Behavior.* New Orleans, LA.

Monk, J.J., & Stallings, W.M. (1970). Effect of item order on test scores. *Journal of Educational Research, 63*, 463-465.

Munz, D.C., & Smouse, A.D. (1968). Interaction effects of item difficulty sequence and achievement anxiety reaction on academic performance. *Journal of Educational Psychology, 59*, 370-374.

Owen, S.V., & Froman, R.D. (1987). What's wrong with three-option multiple choice items? *Educational and Psychological Measurement, 47*, 513-520.

Plake, B.S. (1980). Item arrangement and knowledge of arrangement on test scores. *Journal of Experimental Education, 49*, 56-58.

Plake, B.S., & Ansorge, C.J. (1984). Effects of item arrangement, sex of the subject, and test anxiety on cognitive and self-perception scores in a nonquantitative content area. *Educational and Psychological Measurement, 44*, 423-430.

Plake, B.S., Ansorge, C.J., Parker, C.S., & Lowry, S.R. (1982). Effects of item arrangement, knowledge of arrangement, test anxiety and sex on test performance. *Journal of Educational Measurement, 19*, 49-57.

Plake, B.S., Melican, G.J., Carter, L., & Shaughnessy, M. (1983). Differential performance of males and females on easy to hard item arrangements: Influence on feedback at the item level. *Educational and Psychological Measurement, 43*, 1067-1075.

Plake, B.S., Thompson, P.A., & Lowry, S. (1981). Effect of item arrangement, knowledge of arrangement, and test anxiety on two scoring methods. *Journal of Experimental Education, 49*, 214-219.

Ramos, R.A., & Stern, J. (1973). Item behavior associated with changes in the number of alternatives in multiple choice items. *Journal of Educational Measurement, 10* 305-309.

Sax, G., & Carr, A. (1962). An investigation of response sets on altered parallel forms. *Educational and Psychological Measurement*, 22, 371-376.

Schmitt, J.C., & Scheirer, C.J. (1977). The effect of item order on objective tests. *Teaching of Psychology, 4*, 144-145.

Smouse, A.D., & Munz, D.C. (1968). The effects of anxiety and item difficulty on achievement test scores. *Journal of Psychology, 68*, 181-184.

Spiers, P.A., & Pihl, R.O. (1976). The effect of study habits, personality and order of presentation on success in an open-book objective examination. *Teaching of Psychology*, 3, 33-34.

Spies-Wood, E. (1980). Learned helplessness and item difficulty ordering. *Psychologia Africana, 19*, 29-40.

Straton, R.G., & Catts, R.M. (1980). A comparison of two, three and four-choice items tests given a fixed total number of choices. *Educational and Psychological Measurement, 40*, 357-365.

Towle, N.J., & Merrill, P.F. (1975). Effects of anxiety type and item-difficulty sequencing on mathematics test performance. *Journal of Educational Measurement, 12*, 241-249.