SEC

AD-A211 041 REPORT DOCUMENTATION PAGE

(2)

1a. RESTRICTIVE MARKINGS	
2a. SECURITY CLASSIFICATION AUTHORITY UNCLASSIFIED	3. DISTRIBUTION / AVAILABILITY OF REPORT Approved for public release; distribution unlimited.
2b. DECLASSIFICATION / DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S)	5. MONITORING ORGANIZATION REPORT NUMBER(S) AFOSR-TR- 89-1031
6a. NAME OF PERFORMING ORGANIZATION The Milton S. Hershey Medical Center	6b. OFFICE SYMBOL (If applicable)
6c. ADDRESS (City, State, and ZIP Code) Hershey, PA 17033	7a. NAME OF MONITORING ORGANIZATION AFOSR/NL
	7b. ADDRESS (City, State, and ZIP Code) Bldg 410 Bolling AFB, DC 20332-6448
8a. NAME OF FUNDING / SPONSORING ORGANIZATION Air Force Office of Scientific Research	8b. OFFICE SYMBOL (If applicable) NL
	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER Grant AFOSR-87-0216A
8c. ADDRESS (City, State, and ZIP Code) AFOSR/PKD Building 410 Bolling AFB, DC 20332-6448	10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO.
61102F	2313	A7	

DTIC ELECTE AUG 08 1989 S E

11. TITLE (Include Security Classification)
Slope-Controlled Performance Testing

12. PERSONAL AUTHOR(S)
Marshall B. Jones

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Final Report	FROM 6/15/87 TO 9/30/88	1989, July 18	2

16. SUPPLEMENTARY NOTATION

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

Cognitive ability tests, though promising in other respects, often show pronounced practice effects and have weak test-retest reliabilities. One reason for the low reliabilities appears to be that practice effects themselves vary from individual to individual, so that subjects differ not only in the levels at which they are performing when testing ends but also in the slopes leading up to those levels. Since slope of the performance curve late in practice has been shown to affect performance at reacquisition (retest), uncontrolled variation in slope may lower test-retest reliability. A possible approach to this problem is experimentally to control slope during testing so that all subjects are improving at roughly the same rates when testing ends. Under this treatment testing (practice) is continued until an individual's improvement from the just-preceding to the last block of trials drops below a critical value; at this point testing stops. Individual subjects vary in both level of performance at the end of testing and number of test blocks, but they are all roughly comparable in the slopes of their performance curves at the end of testing (acquisition).

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT	21. ABSTRACT SECURITY CLASSIFICATION
☐ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	
22a. NAME OF RESPONSIBLE INDIVIDUAL Dr. John F. Tangney	22b. TELEPHONE (Include Area Code) (202) 767-5021 / 22c. OFFICE SYMBOL NL

DD FORM 1473, 84 MAR 83 APR edition may be used until exhausted. All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE

89 8 00 1 12 UNCLASSIFIED 20 JUL 1989

SLOPE-CONTROLLED PERFORMANCE TESTING

INTRODUCTION

Following a suggestion first put forward by Estes (1974), cognitive-ability testing has developed as a new paradigm in differential psychology. In this new paradigm a particular task is first studied experimentally, using latency or error scores as dependent variables, and then modelled mathematically in terms of relevant psychological processes. The parameters in these models vary from individual to individual and, hence, constitute so many measures of individual variation.

In the 15 years that have intervened since Estes' suggestion cognitive-ability testing has developed strongly and shown much promise. It has also, however, been plagued by several technical difficulties, among the most important being the following two. First, the tasks employed have been performance rather than knowledge tasks and have, like most performance tasks, shown practice effects. In a knowledge test the subject does not usually know whether he or she is right or wrong. As a result, practice effects are limited to auxiliary aspects of the task (test-taking skills) and, while they exist, are not large (Messick & Jungblut, 1981; Wing, 1980). In performance testing, however, it is usually not possible to prevent the subject from obtaining some idea as to how well or poorly he or she has done. As a consequence, subjects do better on a test the more times it is administered to them (Bittner et al, 1983; Kennedy et al, 1981). In effect, each test administration is a trial of practice.

Second, cognitive-ability tests tend to have low test-retest reliabilities (Kyllonen, 1986). The reasons are probably several. One is that parameters in theoretical models are usually estimated by the difference between two positively correlated direct measures; such differences, however, include the error terms for both direct measures while excluding overlapping true-score variation and tend, as a consequence, to be unreliable. Another reason is that, where all subjects are given the same number of trials, slope is not controlled. Practice effects, like all other behaviors, vary from individual to individual. As a result, subjects differ not only in the levels at which they are performing when practice ends but also in the slopes leading up to those levels.

This last reason is central to the present work. Jones (1989) recently reported two experiments in which all subjects practiced one or another motor-skills task for a fixed number of sessions; there then followed a no-practice interval, after which all subjects resumed practice under the same conditions as obtained in acquisition. The results in both experiments were the same. The flatter an individual's performance curve late in practice and the earlier it became flat or nearly so, the better that individual performed at reacquisition. Since level of performance at the end of acquisition was statistically controlled, the effect of slope and that of level were independent.

Consider, for example, two subjects both of whom are performing at the same level after k trials of practice (when acquisition ends) but one of whom, A, has been performing at or near this level for several trials while the other, B, has only just arrived there after a rapid improvement late in practice. Subject A will perform better at reacquisition than subject B.

Why? Basically because A has practiced the response to be retained over the no-practice interval more than B has. The latter may have had only a single trial of the response to be retained, whereas A has been practicing that response for "several trials."

The application to test-retest reliability is direct. If all subjects are given the same number of test administrations and practice effects (slopes) vary from individual to individual, as inevitably they will, then these variations in slope will generate differences in performance at retest that register as unexplained variance and, therefore, lower test-retest reliability. In the present work individual slopes are controlled or, more accurately, held within a fixed range of variation, thereby hopefully eliminating one source of unreliability between test and retest.

To find out if the results Jones obtained in motor-skill testing apply to performance testing, two experiments were conducted. The first was a preliminary experiment to select appropriate tests for the main experiment and to fix suitable "stop" values for the slope-control condition. The second or main experiment tested the hypothesis that slope control improves test-retest reliability.

PRELIMINARY EXPERIMENT

Subjects

The subjects were 511 basic airmen, who were tested at the Human Resources Laboratory (HRL) at Brooks Air Force Base, Texas.

Tests

Each subject was administered seven tests: Physical Identification, Name Identification, Meaning Identification, Memory Search, Sentence

Verification, Nonverbal Arrow Test, and HiLo Matching for Meaning and Position. These tests were selected in consultation with HRL personnel from the library of computer-administered tests developed as part of the Learning Abilities Measurement Program (LAMP). The criteria used in their selection were: (a) diversity of information-processing function and (b) practical importance to the Air Force as judged by the HRL personnel. The seven tests are briefly described as follows:

Physical Identification. The subject is required to report as quickly as possible whether or not two symbols appearing simultaneously on the screen are identical.

Name Identification. The subject is required to report as quickly as possible whether or not two letters appearing simultaneously on the screen have the same name. For example, A and a have the same name but A and B do not.

Meaning Identification. The subject is required to report as quickly as possible whether or not two words appearing simultaneously on the screen have the same meaning.

Memory Search. The subject is presented with a set of symbols. The display is removed and after a short interval the subject is presented with a single symbol and asked to indicate whether or not it was a member of the original set.

Sentence Verification. The subject is asked to indicate whether a sentence such as "A precedes B" or "A is followed by B" is consistent with an arrangement of the letters A and B such as "AB" or "BA".

The Arrow Test. This test is the same as Sentence Verification except that instead of words such as "precedes" or "follows" the subject is presented with arrows:

A B, A B, A B, or A B,

where the slash indicates negation.

HiLo Matching. The subject is presented with a 2x2 matrix in which one of the squares on the left contains either an X or an O and one of the squares on the right contains either the word "Hi" or the word "Lo." A response is correct if the row indicated by an X (or not indicated by an O) is correctly identified as Hi or Lo.

The dependent measures for all tests were: percent correct (PCA), response time in milliseconds on all trials (CTA), and response time in milliseconds on correct trials (CTR). CTA and CTR for a block of trials were calculated by first taking the median in each set of eight trials and then averaging the four medians. Individual trials on all tests lasted between 4 and 8 seconds.

Procedure

All subjects were administered four blocks of 32 trials on all tests, followed after approximately 1 hour by two more 32-trial blocks. Order of testing was counterbalanced in a 7x7 Latin square design.

Results

Three tests (Name Identification, Meaning Identification, and Memory Search) showed no appreciable practice effects on either PCA or CTR. Since more or less sustained improvement with practice is a sine qua non for slope control, these three tests were excluded from further consideration.

The remaining four tests organized themselves into a 2x2 pattern. Two tests (HiLo Matching and Sentence Verification) showed practice effects primarily on PCA, while the other two tests (the Arrow Test and Physical Identification) showed practice effects primarily on CTR. Further, two of the tests (HiLo Matching and the Arrow Test) showed moderate practice effects, while the other two (Sentence Verification and Physical Identification) showed weak practice effects.

Table 1 presents results for these four tests. The first six rows contain the block means for the four tests, in percent for HiLo and Sentence Verification and in seconds for the Arrow Test and Physical Identification. The next row gives the average standard deviation over the first four blocks for each test. Effect size (d), the next row, was calculated as

for HiLo and Sentence Verification and the same except with reversed sign, that is, $(X_1 - X_4)$, for the Arrow Test and Physical Identification. Finally, the last row in Table 1 presents the average correlation between the first four and the last two blocks (eight correlations). As can be seen, the effect sizes fall into two pairs: one at .80 (HiLo Matching and the Arrow Test) and the other at .40 (Sentence Verification and Physical Identification). Further, the average reliabilities are modest, allowing much room for improvement.

TABLE 1

Means by Block (32 Trials), Average Standard Deviation, Practice Effect Size, and Average Test-Retest Reliability for HiLo Matching, Sentence Verification, the Arrow Test, and Physical Identification.

	Test[1]			
Item	HiLo Matching	Sentence Verification	Arrow Test	Physical Identification
x, Block 1	78.1	72.9	2.21	0.534
x, Block 2	84.5	75.7	1.93	0.510
x, Block 3	87.4	77.6	1.73	0.514
x, Block 4	89.0	78.0	1.66	0.490
x, Block 5	89.0	81.2	1.58	0.519
x, Block 6	91.3	80.9	1.49	0.526
SD, Average	13.0	15.4	0.69	0.103
d	0.84	0.33	0.80	0.43
r	0.54	0.60	0.62	0.45

[1] The dependent measure for HiLo Matching and Sentence Verification is PCA (percent) and for the Arrow Test and Physical Identification CTR (seconds).

Stop Regions

The next step is to determine slope values below which testing stops for a given individual and test. In the main experiment the procedure will be for all subjects to take two block of 32 trials. If improvement from the first to the second block (properly oriented mean difference) is less than or equal to the stop value, testing stops for that subject. If improvement from the first to the second block exceeds the stop value, the subject receives a third block of 32 trials. If improvement from the second to the third block is less than or equal to the stop value, the subject receives no more trials;

otherwise, testing continues into a fourth block and then stops regardless of how much or how little improvement the subject makes. In all, therefore, there are three stopping points, after two, three, and four blocks of testing. Some subjects, moreover, will continue to improve from the third to the fourth block at a rate exceeding the stop value. Any such subject will be said to have "escaped" slope control.

In fixing stop values three principles were observed:

--The stop values may be zero or positive but not negative;

--The proportion of subjects who escape control should not be larger than 10%;

--The number of subjects at the three stopping points should be as nearly equal as possible consistent with the first two principles.

Applying these principles, one gets the results which appear in Table 2. The first row gives the "stop regions" for each test. A stop region is defined by all values of improvement less than or equal to the stop value. For HiLo Matching and Sentence Verification the stop value is 0. If from any block to the next a subject responds correctly on the same or a smaller number of trials, testing stops. For the two response-time tests the stop values are .13 and .01 seconds. If from any block to the next CTR drops by .13 or .01 seconds or less for the Arrow Test or Physical Identification respectively, testing stops.

TABLE 2

Stop Regions and Stop Numbers for HiLo Matching, Sentence Verification, Arrow Test, and Physical Identification.

	Test			
Item	HiLo Matching	Sentence Verification	Arrow Test	Physical Identification
Stop Region	0	0	.13	.01
N (Stop 2)	178	231	177	178
N (Stop 3)	176	176	172	240
N (Stop 4)	157	101	158	90
N (Total)	511	508	507	508
N (Escape)	47	29	48	42

The next three rows give the stop numbers for the three stopping points that would have obtained in the preliminary study if the stop regions indicated in the first row had been applied. The next row gives the total number of subjects who provided valid data on all 24 8-trial sets for that measure and test. The last row gives the numbers of subjects who would have escaped. The main points are that the escape percentage is held under 10% for all tests and that the numbers at the three stopping points are not greatly imbalanced.

MAIN EXPERIMENT

Subjects

The subjects were 347 basic airmen at Brooks Air Force Base, Texas.

Tests and Procedures

The tests were HiLo Matching, Sentence Verification, the Arrow Test, and Physical Identification. The stop values were those indicated by the preliminary experiment. The design is a Latin square with four test-treatment groups, where each group is administered all four tests and each group also takes one test under each of the four treatment conditions. Thus, each group is administered only one test under the slope-controlled condition and that one is different in each of the four groups. Retesting is the same for all four groups and consists of two blocks of testing on all tests. Order of testing is counterbalanced within each of the four groups in a 4x4 Latin square. Altogether, therefore, there are 16 groups of subjects.

Results

Tables 3-6 present sample sizes, means, and standard deviations for the four tests. The results conform closely to what would have been expected from the preliminary experiment.

TABLE 3

Sample Sizes, Means, and Standard Deviations for the HiLo Matching Test.

Measure	Test/ Retest	Block	Treatment: Final Two	Final Three	Final Four	Slope Control
N			83	88	87	87
X^1	Test	1	80.8	78.3	77.9	75.4
		2	86.4	84.0	83.7	80.0
		3		86.2	84.5	
		4			86.0	
		X^2				82.3
	Retest	1	91.9	90.3	93.3	90.2
		2	93.6	91.3	92.6	92.1
SD^1	Test	1	13.2	14.9	16.7	13.1
		2	12.6	14.5	14.7	13.3
		3		12.7	15.4	
		4			14.6	
		X^2				13.4
	Retest	1	8.1	11.4	7.5	8.5
		2	6.9	11.8	6.7	8.3

[1] Means and standard deviations are in percentages.

[2] "X" refers to the last block of trials a subject took under the slope-control condition.

TABLE 4

Sample Sizes, Means, and Standard Deviations for the Sentence Verification Test.

Measure	Test/ Retest	Block	Treatment: Fixed Two	Fixed Three	Fixed Four	Slope Control
N			87	84	88	87
X[1]	Test	1	72.0	70.7	70.6	75.2
		2	73.4	70.9	75.0	80.0
		3		74.6	78.4	
		4			76.7	
		X[2]				77.4
	Retest	1	79.7	80.4	81.2	85.4
		2	78.9	81.9	80.7	85.5
SD[1]	Test	1	16.3	17.9	17.4	17.0
		2	17.1	18.1	17.1	14.3
		3		17.5	17.1	
		4			16.2	
		X[2]				15.5
	Retest	1	15.1	16.1	14.8	13.8
		2	17.0	14.8	17.3	13.5

[1] Means and standard deviations are in percentages.

[2] "X" refers to the last block of trials a subject took under the slope-control condition.

TABLE 4

Sample Sizes, Means, and Standard Deviations for the Sentence Verification Test.

Measure	Test/ Retest	Block	Treatment: Fixed Two	Fixed Three	Fixed Four	Slope Control
N			87	84	88	87
X[1]	Test	1	72.0	70.7	70.6	75.2
		2	73.4	70.9	75.0	80.0
		3		74.6	78.4	
		4			76.7	
		X[2]				77.4
	Retest	1	79.7	80.4	81.2	85.4
		2	78.9	81.9	80.7	85.5
SD[1]	Test	1	16.3	17.9	17.4	17.0
		2	17.1	18.1	17.1	14.3
		3		17.5	17.1	
		4			16.2	
		X[2]				15.5
	Retest	1	15.1	16.1	14.8	13.8
		2	17.0	14.8	17.3	13.5

[1] Means and standard deviations are in percentages.

[2] "X" refers to the last block of trials a subject took under the slope-control condition.

TABLE 6

Sample Sizes, Means, and Standard Deviations on Physical Identification

Measure	Test/ Retest	Block	Treatment: Fixed Two	Fixed Three	Fixed Four	Slope Control
N			88	87	87	82
X^1	Test	1	537	540	540	571
		2	475	486	504	494
		3		507	483	
		4			500	
		X^2				506
	Retest	1	488	502	492	498
		2	472	491	481	480
SD^1	Test	1	85	93	103	222
		2	67	78	90	105
		3		109	71	
		4			121	
		X^2				138
	Retest	1	72	87	80	84
		2	66	98	74	81

[1] Means and standard deviations are in milliseconds.

[2] "X" refers to the last block of trials a subject took under the slope-control condition.

Tables 7-10 present correlational results for the four tests. Table 7, for example, concerns HiLo Matching. The third column is the most important. It contains the correlations (reliabilities) between the last block of trials in testing and the first block in retesting. For those subjects who received two (three or four) blocks of trials in testing, the correlation is between the second (third or fourth) block in testing and the first block in retesting. For the slope-control condition, reliability is the correlation between the last block of trials a subject received in testing (which could be his second, third, or fourth block) and the first block in retesting. The reliability for the slope-control condition is 0.568, larger than one but smaller than two other reliabilities.

TABLE 7

Correlational Results for HiLo Matching.

Treatment	Test Blk 1-Blk 2	Retest Blk 1-Blk 2	Reliability Attenuated	Reliability Unattenuated
Fixed 2	.670	.734	.701	1.000
Fixed 3	.753	.872	.702	0.866
Fixed 4	.834	.626	.537	0.743
Slope Control	.765	.647	.568	0.807

TABLE 8

Correlational Results for Sentence Verification.

Treatment	Test Blk 1-Blk 2	Retest Blk 1-Blk 2	Reliability Attenu-ated	Unattenu-ated
Fixed 2	.752	.804	.754	0.970
Fixed 3	.821	.792	.661	0.820
Fixed 4	.680	.815	.812	1.091
Slope Control	.748	.786	.752	0.981

TABLE 9

Correlational Results for the Arrow Test.

Treatment	Test Blk 1-Blk 2	Retest Blk 1-Blk 2	Reliability Attenu-ated	Unattenu-ated
Fixed 2	.840	.907	.783	0.897
Fixed 3	.759	.686	.736	1.020
Fixed 4	.856	.825	.805	0.958
Slope Control	.739	.888	.765	0.944

TABLE 10

Correlational Results for Physical Identification.

Treatment	Test Blk 1-Blk 2	Retest Blk 1-Blk 2	Reliability Attenuated	Reliability Unattenuated
Fixed 2	.844	.810	.719	0.870
Fixed 3	.832	.787	.653	0.807
Fixed 4	.878	.872	.691	0.789
Slope Control	.908	.842	.502	0.574

With sample sizes on the order of 80-90 correlational level varies considerably. It could be, therefore, that the poor result for slope control is due to that group's happening to have poor correlations on that test in general. Since all groups received two blocks of testing initially and two blocks at retest, this possibility can be checked by prorating reliability against the correlations obtained in these two pairs of blocks. The fourth column in Table 7 was obtained by dividing the reliability for a given treatment by the geometric mean of the correlations between Block 1 and Block 2 in testing and Block 1 and Block 2 in retesting. The result may be understood as a sort of "unattenuated" correlation. So corrected, slope control still ranks third among the four treatment conditions.

Tables 8-10 are laid out in the same way as Table 7 and show the same result. Slope control (attenuated) ranks third twice and fourth in the three tables. Unattenuated it ranks second, third, and fourth. No matter how one

looks at them, the results are emphatically negative. In this experiment at least, slope control does not improve test-retest reliability.

Discussion

Jones' result regarding slope as a predictor of performance at reacquisition was obtained with tasks and procedures usual for motor-skill studies. These procedures differ in several key respects from those used in the present study or in performance testing generally. Three differences are especially clear. First, a single data point in Jones' study was taken from a session of testing that typically lasted approximately 15 minutes. A single data point in the present study was taken from a block of 32 trials, which typically lasted approximately 4 minutes. There was, therefore, roughly a fourfold difference in the amount of testing time represented by a single data point. Second, the test-retest interval in the present study was approximately 1 hour, whereas in Jones' original study it varied between 4 and 18 months. Third, acquisition testing in Jones' original study was distributed, with usually more than a day between test sessions, whereas in the present experiment practice was massed; all acquisition testing (or retesting, for that matter) took place in a single sitting.

If the amount of testing time per data point had been short enough to make the results unreliable, the fact would certainly help to explain the negative outcome. There is no evidence, however, that testing time in the present study was that short. Correlational levels were somewhat lower than in Jones' original study but not enough so to account for the complete absence of an effect that was both strong and consistent in Jones' original study.

The difference in retest interval is, of course, enormous; but it would be easier to see a role for it in explaining the negative result if it had been the other way around, that is, if the present study had used the long interval and Jones' original study the short one. Lengthening the retest interval might be expected to attenuate an effect to the point of eliminating it; but shortening the retest interval would not seem likely to do so. Again, therefore, while a major difference unquestionably exists, it does not offer a ready explanation for why the present experiment turned out so emphatically negative.

The difference in distribution of practice offers a possible explanation. Many effects take place within a practice session that do not play an important role between sessions. All four of the test blocks in the present experiment lasted approximately as long as one test session in Jones' original study. Within a 15-minute test session, however, fatigue, loss of concentration, even boredom can become major factors. Hence, when one compares two points within a session any difference is likely to reflect fatigue, loss of concentration, or boredom. Between sessions, however, the same factors play little or no role, not because they aren't present but just the contrary, because they are present in roughly the same degree in both sessions. As a result, differences between sessions tend to reflect differences in skill acquisition primarily.

If the above account of why the present experiment turned out negative is correct, then experimental slope control would not seem to have a future in performance testing. Performance testing for purposes of prediction, selection, or assignment is universally done today in a single sitting. The subject is administered many trials but all in a single session. Retesting

is usually done in a separate session, possibly after a retest interval lasting months. It would be technically possible to distribute original testing over several separated blocks of testing within a single session. It might even be possible to carry out original testing in a series of separate sessions. The likelihood, however, that any such testing schedule will be implemented is remote. If the failure of slope control in the present experiment is due to massing practice in a single test session, then that failure will generalize beyond the four tests and particular procedures used in the present study, because all performance testing for personnel purposes is carried out at present in a single test session. It follows that experimental slope control is not a feasible way of improving the test-retest reliability of cognitive tests.

REFERENCES

Bittner, A.C., Jr., Carter, R.C., Krause, M., & Harbeson, M.M. (1983). Performance Evaluation Tests for Environmental Reserch (PETER): Moran and computer batteries. Aviation, Space, and Environmental Medicine, 54, 923-928.

Estes, W.K. (1974). Learning theory and intelligence. American Psychologist. 54, 740-749.

Jones, M.B. (1989). Individual differences in skill retention. American Journal of Psychology, 102, 183-196.

Kennedy, R.S., Bittner, A.C., Jr., Carter, R.C., Krause, M., Harbeson, M.M., McCafferty, D.B., Pepper, R.L., & Wiker, S.F. (1981). Performance Evaluation Tests for Environmental Research (PETER): Collected papers (NBDL-80R008). New Orleans, LA: naval Biodynamics Laboratory.

Kyllonen, P.C. (1985). Theory-based cognitive assessment (ASFHRL-TP-85-30). Brooks Air Force Base, TX: Air Force Human Resources Laboratory.

Messick, S., & Jungblut, A. (1981). Time and method in coaching for the SAT. Psychological Bulletin, 89, 191-216.

Wing, H. (1980). Practice effects with traditional test items. Applied Psychological Measurement, 4, 141-155.