DOCUMENT RESUME

ED 352 391 TM 019 295

AUTHOR Roos, Linda L.; And Others
TITLE The Effects of Feedback in Computerized Adaptive and Self-Adapted Tests.
PUB DATE Apr 92
NOTE 23p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 21-23, 1992).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Ability Identification; *Adaptive Testing; Algebra; Algorithms; Comparative Testing; *Computer Assisted Testing; Difficulty Level; *Feedback; *Graduate Students; Higher Education; Statistics; Test Anxiety; *Test Items; *Undergraduate Students

ABSTRACT

Computerized adaptive (CA) testing uses an algorithm to match examinee ability to item difficulty, while self-adapted (SA) testing allows the examinee to choose the difficulty of his or her items. Research comparing SA and CA testing has shown that examinees experience lower anxiety and improved performance with SA testing. All previous research concerning SA testing has presented item feedback to the examinee before asking the examinee to choose the next item difficulty level. Moreover, item feedback has typically not been presented to examinees in previous CA testing research. The effects of presenting, versus withholding, item feedback in SA tests were studied for 135 graduate and 228 undergraduate students (128 males and 235 females). The instrument was a computerized algebra test to assess skills needed for a statistics class. Examinees administered the SA tests tended to obtain significantly higher ability estimates than did those who were administered the CA tests. Also, those taking the SA tests reported significantly lower post-test state anxiety than did those taking the CA tests. Interaction between test type and feedback was not found, suggesting that examinees are able to use the implicit feedback they receive when answering items. Five tables present study findings. (SLD)

The Effects of Feedback in Computerized Adaptive and Self-Adapted Tests

Linda L. Roos

Barbara S. Plake

Steven L. Wise

University of Nebraska–Lincoln

Paper presented at the 1992 annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Abstract

Computerized adaptive (CA) testing uses an algorithm to match examinee ability to item difficulty while self-adapted (SA) testing allows the examinee to choose the difficulty of his/her items. Research comparing SA testing and CA testing has shown that examinees experience lower anxiety and improved performance with SA testing. All previous research concerning SA testing has presented item feedback to the examinee before asking the examinee to choose the next item difficulty level. Moreover, item feedback has typically not been presented to examinees in previous CA testing research. This study looked at the effects of presenting, versus withholding, item feedback in SA tests. Additionally, previous research comparing SA and CA tests was extended.

The Effects of Feedback in Computerized Adaptive and Self-Adapted Tests

Introduction

The advent of item response theory (IRT) allows examinee test performance to be compared using the same scale regardless of which items from a unidimensional item pool are administered to examinees. Therefore, under the tenets of IRT, examinee ability estimation is independent of the set of items administered from a unidimensional pool of calibrated items. Computerized adaptive (CA) testing, an application of IRT, employs a computer algorithm that matches item difficulty to examinee ability level. The algorithm's selection of the next item to be administered is based on the examinee's responses to previously administered items. A variant of CA testing, self-adapted (SA) testing, was proposed by Rocklin and O'Donnell (1987). Self-adapted testing allows examinees to choose the difficulty levels of the items administered.

Rocklin and O'Donnell compared examinee performance on an SA test with the performances of examinees taking two conventional computerized tests from the same 40-item pool. One of the conventional computerized tests consisted of the 20 most difficult items, while the other consisted of the 20 easiest items. Rocklin and O'Donnell found that the examinees who were administered an SA test obtained significantly higher ability estimates than examinees administered either of the conventional computerized tests. Additionally, Rocklin and O'Donnell point out that the difference between SA tests and CA tests lies in the fact that a CA test is tailored only to an examinee's estimated ability level while an SA test is tailored to the examinee's perceived ability level taking into consideration current motivational and affective characteristics.

Wise, Plake, Johnson, & Roos (1991) found that examinees who were administered an SA test obtained a significantly higher mean ability score than those administered a CA test. Examinees who were administered the SA test also reported significantly lower mean post-test state anxiety than the examinees who were administered the CA test. Wise, et al. (1991) reported that those examinees who took the SA test also took significantly longer to complete the test and had a significantly larger standard error of ability.

A basic assumption of SA testing is that an examinee requires explicit item feedback in order to make intelligent item level choices on subsequent items. To this end, previous investigations of SA testing have always presented some type of item feedback to the examinee (Rocklin & O'Donnell (1987); Wise, et al. (1991); Johnson, Roos, Wise & Plake (1991)). Item feedback has not been presented to examinees in most studies of CA testing research.

One factor that has been shown to influence motivational and affective characteristics of examinees is item feedback. Research has shown mixed results in terms of effects of feedback on performance and anxiety level. Betz (1977) reports higher test performance for those examinees who receive feedback. Gialluca and Weiss (1980) report that feedback has no significant effect on examinee performance. Prestwood and Weiss (1978) found that anxiety was not significantly higher for those examinees who received feedback than for those who did not. Gilmer (1979) concluded that feedback increased anxiety, especially for low-ability examinees. Rocklin and Thompson (1985) found that, in general, performance was improved by feedback especially for the examinees administered an easy test. They also found that low anxious examinees performed better on average on a hard test than they did on an easier test while the opposite was true of moderately anxious students.

These mixed results lead to questions about the effects of feedback on test performance and anxiety in both SA and CA testing. It is difficult to ascertain whether the reported gains realized by SA tests in terms of higher performance and less anxiety are the result of the type of test or the feedback. Rocklin and O'Donnell (1987) noted that, in SA testing, "an examinee has access to a variety of information (including current affective and motivational states) relevant to optimal item selection" (p. 318). Feedback is clearly a major piece of information available to an examinee in SA testing. In this study, we were interested in comparing the effects of having, versus not having, item feedback in SA testing. If explicit item feedback is necessary for examinees to make effective item choices, then the differences between SA and CA tests in terms of examinee test performance should not be found in the absence of item feedback. That is, the importance of feedback should be shown through an interaction between type of test and the presence or absence of feedback.

Method

Subjects

The subjects were 363 students enrolled in introductory statistics classes at a large midwestern university during the summer and fall of 1991. The subjects included about one-third graduate (135) and about two-thirds undergraduate (228) students. There were 128 (35.3%) males and 235 (64.7%) females.

Instruments

The primary instrument used in this study was a computerized algebra test designed to assess whether students possess the algebra skills necessary for successful completion of an introductory statistics course. The test items utilize a four-option multiple choice format and each examinee was

administered 20 items. The items were chosen from a pool of 91 items testing basic algebra skills. The pool of 91 items was calibrated using a modified one-parameter IRT model in which the lower asymptote of each item characteristic curve was fixed at .20. Model fit was found acceptable using Yen's Q_1 statistic. Wise, et al. (1991) provide a detailed explanation of the development of the item pool. Four versions of the test were administered—SA with and without feedback and CA with and without feedback. Item feedback was given by indicating the correct answer after each question.

The tests were administered using IBM PS/2 Model 55SX microcomputers and Microcat™ software. After the algebra test was administered, several questions were administered electronically which were designed to assess examinees' opinions about the type of test they had received.

The CA test used a maximum likelihood algorithm to determine, based on item information, which item should be administered to the examinee considering the examinee's performance on previously administered items. In general, an examinee was given an easier item after answering incorrectly and a more difficult item after answering correctly. Each version of the CA test terminated when 20 items had been administered. The SA test allowed examinees to choose the difficulty level of each item administered. The 91 items were divided into six difficulty levels each containing 15 or 16 items based on the difficulty (b parameter) of each item. The items within each difficulty level were randomly ordered and all examinees received the items in the same order within each difficulty level. After answering an algebra test item, the examinees were asked to choose the difficulty level of the next item. Since no level contained more than 16 items, examinees sometimes exhausted the items from a particular level before

completing the test. When this was the case, examinees were directed to choose an item from another level until 20 items had been administered.

In addition to the algebra test, four other instruments were used. Each used a paper and pencil format. A scale developed by Wise, Johnson, Plake, and Nebelsick-Gullet (1990) was used to measure examinee preferences in test taking. The Revised Mathematics Anxiety Rating Scale (RMARS; Plake & Parker, 1982) was used to measure examinee mathematics anxiety. The Test Anxiety Inventory (TAI; Spielberger, 1980) measured examinee test-taking anxiety. The State Anxiety Scale (Spielberger, Gorsuch, & Lushene, 1970) was administered immediately before and after the algebra test to measure situation-specific anxiety of the examinees.

Procedure

During the first class session, students supplied demographic information, completed the preference scale, the RMARS and the TAI, and signed up for an algebra test administration time. The students were informed that those who did not score above a particular unspecified cutoff on the algebra test would be required to attend a one hour algebra remediation session to be held early in the term. The students were informed electronically at the end of the testing session if they were required to attend remediation.

Testing was completed during the first two days of the summer classes and during the first week of the fall class. The algebra test was administered in a room containing 12 IBM PS/2 Model 55 microcomputers running Microcat™ software. When students arrived for testing, they were randomly assigned to one of the four test conditions by self-selecting a computer. The four conditions were randomly assigned to the 12 microcomputers throughout the entire testing period. The examinees were first asked to

complete the State Anxiety Scale. Then, each examinee was given a few basic instructions concerning the type of test being administered and he/she started the algebra test. Scratch paper and pencils were provided and the use of calculators was not allowed. No time limit was imposed during testing. An IRT ability score was computed for each examinee using maximum-likelihood estimation. This score was compared to a cutoff value of -.20 to determine those students requiring algebra remediation. The cutoff score was obtained using results of previous studies (Wise, et al. (1991); Johnson, et al. (1991)). Upon completion of the algebra test, the examinees were asked to again complete the State Anxiety Scale. The examinees then answered questions concerning attitudes toward the type of testing they had received. Subsequently, they were informed whether they were required to attend a remediation session.

Data Analysis

Since the purpose of this study was to replicate and extend the results of Wise, et al. (1991), the same four dependent variables were investigated. These included: (a) estimated ability, (b) post-test state anxiety, (c) total testing time, and (d) standard error of estimated ability. The independent variables were test type and feedback resulting in the following four conditions: SA with feedback (SAF), SA without feedback (SANF), CA with feedback (CAF) and CA without feedback (CANF). The variable, years since last algebra course (yrsince), was used as a blocking variable in the analysis of estimated ability. The three blocks used included: (a) less than three years, (b) three to five years, and (c) more than five years. The variable, pre-test state anxiety, was used as a blocking variable in the analysis of post-test state anxiety. The three blocks used were: (a) less than 33 (Low), (b) 33-41 (Medium), and (c)

greater than 41 (High). Three-factor analysis of variance (ANOVA) was used in the analyses involving estimated ability and post-test state anxiety.

Results

Table 1 shows the means and standard deviations for estimated ability broken down by experimental condition and years since last algebra course.

Insert Tables 1 and 2 about here

The results for the ANOVA are shown in Table 2. A significant main effect for test type was found with those examinees who were administered the SA tests obtaining a higher average ability estimate than those who were administered the CA tests. Although the feedback main effect was nonsignificant, feedback did show a significant interaction with yrsince. As a follow-up to the significant interaction, simple main effects tests of feedback at each level of yrsince were performed. The results of these tests are also shown in Table 2. For those examinees whose last algebra course was three to five years ago, there was a significant difference between those examinees who received feedback and those who did not receive feedback, with those examinees who received feedback obtaining a higher average ability estimate.

Table 3 shows the means and standard deviations for post-test state anxiety broken down by experimental condition and pre-test state anxiety. The ANOVA results are shown in Table 4. A significant main effect for test type was found with those who were administered the SA tests on average reporting significantly lower post-test state anxiety than those who were administered the CA tests. None of the interactions were found to be significant.

Insert Tables 3 and 4 about here

Table 5 shows the descriptive statistics for testing time and standard error of ability by each test condition. Because those distributions are quite skewed median values are reported. The median testing times for the SA tests were greater than those for the CA tests; the median testing times for tests in which feedback was given differed by about three and a half minutes while the times for the tests in which no feedback was given differed by about one minute. The median standard error of ability is the same for the SA tests whether or not feedback is given and it is greater than that reported for the CA tests with the CA test without feedback having the smallest error estimate.

Insert Table 5 about here

Discussion

The results of this study were consistent with those found by Wise, et al. (1991). Examinees who were administered the SA tests tended to obtain significantly higher ability estimates than those who were administered the CA tests. Also, those examinees taking the SA tests reported significantly lower mean post-test state anxiety than those taking the CA tests.

The median testing times were longer for the SA tests than the CA tests. Median testing times for examinees who were administered the SA test with feedback were about three and a half minutes longer than for their counterparts taking the CA test. For examinees who were administered the tests without feedback, the median testing times differed by about a minute

with the CA test without feedback taking the least time. Since the examinees taking the SA tests must spend time choosing the difficulty level of each item, this finding is logical. The median standard error of ability was less for examinees taking the CA tests than for the SA tests. The median standard error of ability was the same for the SA tests whether or not feedback was given and it was very similar for both CA tests. The obvious reason for this finding is that the algorithm used by the CA tests is choosing items that will minimize the standard error of ability.

Additionally, the interaction between feedback and years since last algebra course is of interest. For those examinees whose last algebra course was three to five years ago, there was a significant difference between receiving feedback and not receiving feedback with those examinees who received feedback obtaining a higher estimated ability. For those examinees whose last algebra course was less than three years ago or more than five years ago, there was no significant difference between receiving and not receiving feedback. It seems possible that for examinees whose last algebra course was three to five years ago, the feedback was confirmation that they remembered the necessary algebra concepts and that positive reinforcement gave them more confidence on subsequent items. It seems possible that for examinees whose last course was less than three years ago or more than five years ago, explicit feedback did not give them meaningful information about their item performance.

The results of this study indicate the same trade-off outlined in Wise, et al. (1991). The SA test requires more time. Examinees who are administered the SA test obtained a significantly higher mean ability estimate than those who were administered the CA test. Post-test anxiety is lower for those administered the SA test than for those who took the CA test. The

greatest difference in median testing time was about three and a half minutes. SA testing offers the positives of higher mean ability estimates and lower test anxiety in exchange for a small additional amount of testing time.

It is of particular interest that the interaction between test type and feedback was not found. This suggests that explicit feedback is not necessary for SA testing to be beneficial as previous research suggested. It appears that examinees are able to rely on the implicit feedback they receive when answering items. Examinees can judge the difficulty of an item and how likely they were to pass the item without being explicitly informed. The trade-off mentioned previously is less of an issue when SA testing is used without feedback. A reduction in testing time required for SA tests could be realized by not providing feedback while, at the same time, maintaining the positives of higher ability estimates and lower test anxiety in SA tests. Therefore, it appears that the differences between SA and CA tests found in this study and in previous studies do not appear be a function of the presence or absence of explicit feedback. More research, however, into the differences in SA and CA tests is warranted.

Conclusions

This study has implications for the future of CA testing. It is important to better understand the implications of feedback in computer testing. If future studies again show that SA testing results in lowered anxiety levels and increased test performance, then it could prove to be an important alternative to CA testing. The results of the present study concerning feedback have implications for the consideration of feedback in future test designs.

References

Betz, N. E. (1977). Effects of immediate knowledge of results and adaptive testing on ability test performance. Applied Psychological Measurement 1(2), 259-266.

Gialluca, K. A., & Weiss, D. J. (1980). Effects of immediate knowledge of results on achievement test performance and test dimensionality. (Research Report 80-1). Minneapolis, MN: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory.

Gilmer, J. S. (1979). The effects of immediate feedback versus traditional no-feedback in a testing situation. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Johnson, P. L., Roos, L. L., Wise, S. L., & Plake, B. S. (1991). Correlates of examinee item choice behavior in self-adapted testing. Mid-Western Educational Researcher, 4(3), 25-29.

Plake, B S., & Parker, C. S. (1982). The development and validation of a revised version of the mathematics anxiety rating scale. Educational and Psychological Measurement, 42, 551-557.

Prestwood, J. S., & Weiss, D. J. (1978). The effects of knowledge of results and test difficulty on ability test performance and psychological reactions to testing. (Research Report 78-2). Minneapolis, MN: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Rocklin, T. (1989, March). Individual differences in item selection in computerized self adapted testing. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Rocklin, T., & O'Donnell, A. M. (1987). Self-adapted testing: A performance-improving variant of computerized adaptive testing. Journal of Educational Psychology, 79(3), 315-319.

Rocklin, T., & Thompson, J. M. (1985). Interactive effects of test anxiety, test difficulty, and feedback. Journal of Educational Psychology, 77(3), 368-372.

Spielberger, C. D. (1980). Preliminary professional manual for the Test Anxiety Inventory. Palo Alto, CA: Consulting Psychologists Press.

Spielberger, C. D., Gorsuch, R. L., & Lushene, R. E. (1970). Manual for the State-Trait Anxiety Inventory. Palo Alto, CA: Consulting Psychologists Press.

Wise, S. L., Johnson, P. L., Plake, B. S., & Nebelsick-Gullet, L. (1990, April). Providing examinees the opportunity to review items, skip items and change item choices on computerized achievement tests. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Boston, MA.

Wise, S. L., Plake, B. S., Johnson, P. L., & Roos, L. L. (1991, April). A comparison of self-adapted and computerized adaptive tests. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Table 1

Descriptive Statistics for Estimated Ability By Experimental Condition and Years Since Last Algebra Course

Years Since Last Algebra Course	Experimental Condition											
	SAF			SANF			CAF			CANF		
	Mean	SD	n	Mean	SD	n	Mean	SD	n	Mean	SD	n
Less than 3 years	0.64	1.14	20	0.63	1.37	24	0.59	0.72	25	0.32	0.63	24
3 to 5 years	0.66	1.37	24	-0.60	0.93	36	0.22	0.95	34	-0.25	1.26	23
More than 5 years	-0.19	1.13	41	-0.02	1.29	31	-0.40	0.96	32	-0.40	1.11	49
All Examinees	0.25	1.26	85	0.14	1.21	91	0.10	0.98	91	-0.18	1.08	96

Table 2

ANOVA Summary Table for Estimated Ability

Source	SS	df	MS	F	F Prob.
Test Type	4.80	1	4.80	4.05	.045
Feedback	3.43	1	3.43	2.90	.090
Feedback at Less than 3 years	0.42	1	0.42	0.36	.551
Feedback at 3 to 5 years	8.42	1	8.42	7.11	.008
Feedback at More than 5 years	0.04	1	0.04	0.04	.849
Yrsince	37.93	2	18.97	16.01	<.001
Test Type by Feed	0.08	1	0.08	0.07	.796
Test Type by Yrsince	0.23	2	0.12	0.10	.906
Feed by Yrsince	7.59	2	3.79	3.20	.042
Test Type by Feed by Yrsince	0.99	2	0.50	0.42	.658
Within Cell	415.95	351	1.19		
Total	471.10	362			

Table 3

Descriptive Statistics for Post-Test State Anxiety By Experimental Condition and Pre-Test State Anxiety Level

Pre-Test State Anxiety Level	Experimental Condition											
	SAF			SANF			CAF			CANF		
	Mean	SD	n	Mean	SD	n	Mean	SD	n	Mean	SD	n
Low	27.85	6.99	20	31.41	10.28	27	31.79	10.33	28	30.37	8.36	27
Medium	40.82	6.29	22	37.36	8.47	33	41.42	7.61	31	38.46	7.75	28
High	49.21	12.90	43	46.40	9.15	30	52.97	9.04	32	49.72	10.52	40
All Examinees	42.01	13.37	85	38.59	10.99	90	42.52	12.43	91	40.91	12.22	95

Table 4

ANOVA Summary Table for Post-Test State Anxiety

Source	SS	df	MS	F	F Prob.
Test Type	353.87	1	353.87	3.97	.047
Feedback	302.00	1	302.00	3.39	.067
Pre-State Anxiety	22089.13	2	11044.57	123.81	<.001
Test Type by Feedback	44.93	1	44.93	0.50	.478
Test Type by Pre-State Anxiety	125.40	2	62.70	0.70	.496
Feed by Pre-State Anxiety	287.88	2	143.94	1.61	.201
Test Type by Feedback by Pre-State Anxiety	113.23	2	56.617	0.64	.531
Within Cell	31132.76	349	89.206		
Total	54527.99	360	151.47		

Table 5

Descriptive Statistics for Total Testing Time and Standard Error of Ability

Dependent Variable	Experimental Condition	Minimum	Median	Maximum
Testing Time (Minutes)	SAF	9.55	21.63	51.40
	SANF	8.65	19.48	46.60
	CAF	9.32	18.02	43.98
	CANF	9.00	18.41	37.08
Standard Error of Ability	SAF	0.33	0.39	4.27
	SANF	0.32	0.39	18.66
	CAF	0.33	0.36	0.64
	CANF	0.33	0.35	2.08