Effects of Immediate Knowledge of Results and Adaptive Testing on Ability Test Performance

Nancy E. Betz
The Ohio State University

This study investigated the effects of immediate knowledge of results and adaptive testing on performance on a computer-administered test of verbal ability. Examinees were administered either a 50-item conventional test or an adaptive test of verbal ability; half the subjects in each group received immediate knowledge of results (KR) concerning the correctness/incorrectness of each item response, while the other half did not. Subjects within high- and low-ability subgroups were assigned randomly to one of the four resulting experimental conditions. The dependent variable was maximum likelihood ability estimates derived from item response patterns. Results indicated that for the high-ability group, mean test scores under KR conditions were significantly higher than were those under no-KR conditions on both the conventional and adaptive tests. Within the low-ability group, mean test scores were higher under KR conditions than under no-KR conditions, but the difference was statistically significant only within the conventional testing strategy. Low-ability examinees achieved higher average test scores on the adaptive test than on the conventional test, while high-ability examinees performed equally well on the adaptive and conventional tests.

The utility of psychological measurement procedures depends not only on the psychometric characteristics of the measuring instrument but on the "psychological state" of the individual being measured. In the assessment of intelligence and abilities, a critical factor in this "psychological state" is the extent to which an examinee is motivated to do his/her best on the test; motivation is, in fact, one of the *a priori* assumptions underlying the measurement of these variables.

While attempts to maintain examinee motivation at high levels were an integral part of the administration of individual intelligence tests (Terman & Merrill, 1960; Wechsler, 1955), the adequacy of measurement using these tests was limited by the lack of standardization in administration and by the subjectivity in scoring (see Sattler & Theye, 1967 and Weiss & Betz, 1973 for reviews of the literature). Group-administered intelligence and ability tests, while characterized by a high degree of standardization and objectivity, made no provision for ensuring that examinees were motivated to demonstate their full capabilities on the test. Thus, until recently, it was not possible to ensure both optimal psychometric characteristics and maximal motivation of the examinee within the same assessment procedure.

Now, however, computer-assisted testing procedures have made it possible to combine the high degree of standardization and objectivity previously possible only in group-administered

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 1, No. 2 Spring 1977 pp. 259–266

tests with an individualized mode of test administration. These new procedures provide several possible approaches to the maintenance of examinee motivation. Two of these approaches, immediate knowledge of results and "adaptive" (Weiss & Betz, 1973) or "tailored" (Lord, 1970) testing strategies, were implemented in the present study.

Immediate knowledge of results, or KR, has been extensively studied by investigators of human learning, and its facilitative effect on human learning and retention is now considered to be firmly established (e.g., Annett, 1969; Bilodeau & Bilodeau, 1961). Studies of the effects of KR on achievement test performance have yielded conflicting findings; KR was found to increase test scores in studies by Beeson (1973) and Heald (1970) but led to an increased number of errors in studies by Bierbaum (1965), Spencer and Barker (1969), and Strang and Rust (1973).

While several investigators have postulated that immediate KR may increase scores on ability and/or intelligence tests by enhancing examinee motivation (e.g., Bayroff, 1964; Ferguson & Hsu, 1971; Weiss & Betz, 1973), research relevant to this hypothesis is lacking. In studies by Betz (1975) and Sweet and Ringness (1971), examinees obtained higher test scores under KR conditions than under no-KR conditions on a computer-administered test of verbal ability and on the WISC, respectively. A study by Zontine, Richards, and Strang (1972) found no differences in performance on the Peabody Picture Vocabulary Test as a function of KR conditions. In the studies of Betz (1975) and Sweet and Ringness (1971), however, interactions between the effects of KR and racial and socioeconomic status suggested that KR may have facilitative effects on the performance of examinees who have typically been less motivated to do well on ability tests (e.g., lower-class or black examinees), but may not affect the performance of examinees from white, middle-class backgrounds who may in general have higher levels of motivation to do well.

It has also been postulated that adaptive ability testing increases levels of test-taking motivation, thereby increasing test scores, particularly for examinees for whom conventional ability tests are too difficult. Conventional ability tests, constructed to be maximally appropriate in difficulty level for the average ability level in a group of individuals, are of necessity too easy for high-ability examinees and too difficult for low-ability examinees. High-ability examinees may become bored when administered items that are far too easy for them, while low-ability examinees are likely to become frustrated and/or discouraged when confronted with a succession of difficult items. Neither boredom nor frustration or discouragement is conducive to maximal effort and motivation.

Adaptive testing may avoid these problems by administering items selected to be appropriate in difficulty level to each examinee's ability level rather than, as in conventional testing procedures, the mean ability level of a group of examinees. Adaptive tests permit each examinee, regardless of ability level, to respond correctly to about half of the items administered to him/her. Thus, it seems reasonable to hypothesize that adaptive testing may have positive motivating effects on examinees by presenting them with items that are sufficiently difficult to present a challenge yet not so difficult as to seem hopeless. These effects may be particularly apparent for low-ability examinees who, rather than being confronted with a frustrating or even hopeless task, receive items on which there is a moderate probability of success.

Thus, the purposes of the present study were: 1) to examine the effects on performance of immediate knowledge of results and adaptive versus conventional testing strategies on a computer-administered test of verbal ability, and 2) to determine whether the effects of KR and testing strategy differed depending upon the ability level of the examinee.

Method

Subjects

The "high-ability" group consisted of 239 undergraduate students from the introductory

psychology course in the College of Liberal Arts (CLA) at the University of Minnesota. The "low-ability" group consisted of 111 students taking psychology courses in the University's General College (GC).

Admissions standards for CLA are relatively stringent, while GC maintains an "open" admissions policy; the difference between the mean scholastic aptitude test scores in the two groups is highly significant and there is very little overlap between the CLA and GC score distributions on the Minnesota Scholastic Aptitude Test. All 350 tested subjects were volunteers.

Ability Tests

The conventional test consisted of 50 items selected from a pool of about 400 five-alternative multiple-choice vocabulary items normed on a large group of college students. Normal ogive difficulty and discrimination parameters were available for each item; the normal ogive difficulty parameter *(b)* is related to the "proportion correct" index of item difficulty used in traditional test theory; the normal ogive discrimination parameter *(a)* is a function of the item-total score biserial correlation coefficient. The difficulties of the items were concentrated around $b = -.20$ (equivalent to a p-value of .54), and all items had discriminations greater than or equal to $a = .40$ (equivalent to a biserial correlation of .37). The 50 items were administered in the same order to all examinees given the conventional test.

The adaptive test utilized in the present study was the stradaptive testing strategy (Weiss, 1973). To construct the stradaptive test, the 400 items in the original pool were grouped into nine levels, or strata, on the basis of their difficulties. Items ranged in difficulty from $b = +3$ to $b = -3$, and the difficulty range of items within a stratum was .67. There was no overlap in item difficulties between adjacent strata. Within each stratum, the 30 most highly discriminating items available were selected for inclusion in the test; strata at the extreme levels of difficulty did not contain 30 items having minimally acceptable discriminating power (i.e., $a \geq .30$), so these strata consisted of as few as 17 items. Out of the total pool of 400 items, 243 items comprised the final stradaptive test structure.

Examinees were entered into a stratum of the stradaptive test on the basis of their estimated grade-point averages; those examinees reporting high GPAs began the test with more difficult items than did those reporting lower GPAs. Examinees were branched through the stradaptive item structure according to the rule that, following a correct response, the most discriminating item remaining in the next more difficult stratum was administered and, following an incorrect response, the most discriminating item in the next less difficult stratum was administered. Testing was terminated when a ceiling stratum had been identified. Since the items used were five-alternative multiple-choice items, the ceiling stratum was defined as that stratum where the examinee answered 20% or fewer of the items correctly, based on a minimum of five items administered at that stratum. For most examinees, a ceiling stratum was determined after the administration of about 30 items. For a few very high-ability examinees capable of responding above the chance level at even the most difficult stratum, a ceiling stratum could not be identified. In such cases, testing was terminated after the administration of 75 items.

Both the conventional and stradaptive tests were scored using maximum likelihood ability estimation procedures. The scoring formula, based on Birnbaum's three-parameter logistic model, is contained in Birnbaum (1968, Ch. 20, Sec. 20.3) and Samejima (1973, p. 222, Eq. 1.8). Input into the scoring program consisted of the examinee's vector of 1's and 0's, corresponding to correct and incorrect item responses, respectively, and the item difficulty, discrimination, and guessing parameters characterizing each item. The guessing parameter (c) was set at .20, corresponding to the probability of obtaining a correct response through random selection of one of the five possible alternatives for each multiple-choice item.

Procedure

Within the "high-ability" (CLA) and "low-ability" (GC) groups, subjects were assigned at random to one of the four experimental conditions (i.e., the conventional or the stradaptive test administered with or without KR).

All students were tested at individual cathode-ray terminals (CRTs) connected to a Hewlett-Packard 9600E Real-Time computer system. Test items were presented on the CRT screen, and testees indicated their response by typing in the number corresponding to the chosen alternative for each five-alternative multiple-choice item. Instructional screens explaining the operation of the CRTs were provided prior to testing (see DeWitt & Weiss, 1974, pp. 36–53) and a proctor was present in the testing room to provide assistance to any testee having difficulty with the terminal equipment. Students were permitted as much time as necessary to complete the tests and were so informed before testing was begun.

Examinees in the KR conditions received either the message "That's correct" or "That's not correct. The correct answer is x" following each item response. Examinees in the no-KR conditions were immediately administered another item following each item response.

Results

Table 1 shows the results of the three-way analysis of maximum likelihood ability estimates and the means and standard deviations of ability estimates as a function of KR, testing strategy, and subject group.

As indicated in Table 1, there were significant main effects for KR and for subject group. The mean ability estimates obtained by both high- and low-ability subjects were significantly higher when tests were administered under KR conditions than when they were administered under no-KR conditions; the mean ability estimate obtained under KR conditions was -.33, while that under no-KR conditions was -.58. Further, the overall mean level of performance of the high-ability group was significantly higher ($p<.01$) than that of the low-ability group; the mean of the former group over all treatment combinations was -.26, while that of the latter group was -.87.

Although there was no significant effect for testing strategy nor were there significant interaction effects, contrasts on the individual means indicated that both the KR and subject group main effects were moderated by level on the other factors. The means for the eight experimental groups are plotted in Figure 1; the dashed lines in the figure enclose means which are not significantly different from each other but which are significantly different from at least one of the means outside the enclosure.

As shown in Figure 1, KR had substantially more effect on test scores when it was provided in the administration of the conventional test than in the administration of the stradaptive test. On the conventional test, the high-ability-KR mean (-.06) was significantly greater than the high-ability-no-KR mean (-.43), and the low-ability-KR mean (-.87) was significantly higher than the low-ability-no-KR mean (-1.20).

On the stradaptive test, however, while the level of performance of the high-ability group under KR conditions (-.19) was significantly greater than that under no-KR conditions (-.39), the differences for the low-ability group (-.69 and -.72) and for the combined groups (i.e., high-ability and low-ability) were not statistically significant.

Figure 1 also indicates that the size and significance of the group differences in performance were a function of testing conditions. Although the overall level of performance in the high-ability group was significantly higher than that of the low-ability group, the performance levels of low-ability subjects on the stradaptive test and on the conventional test under KR conditions were not significantly lower than those of high-ability subjects taking either test under no-KR conditions. Thus, for testees administered the stradaptive test without knowledge of results, no significant differences were found between the test scores of "high" and "low" ability groups. The performance levels of high-ability subjects were highest under KR conditions, while the

Table 1
Means and Standard Deviations of Maximum Likelihood Ability Estimates for Conventional and Stradaptive Tests in High- and Low-Ability Groups With and Without KR, and Three-Way ANOVA Results

	Experimental Condition						Combined Conditions		
	KR			No-KR					
Test and Group	*N*	Mean	S.D.	*N*	Mean	S.D.	*N*	Mean	S.D.
Conventional Test									
High-Ability	60	-.06	1.04	57	- .43	1.22	117	- .24	1.14
Low-Ability	28	-.87	.84	28	-1.20	1.40	56	-1.03	1.16
Stradaptive Test									
High-Ability	60	-.19	1.21	62	- .39	.91	122	- .29	1.07
Low-Ability	27	-.69	.79	27	- .72	.89	55	- .71	.83
Combined Groups									
Conventional Test	88	-.31	1.05	85	- .68	1.33	173	- .49	1.20
Stradaptive Test	87	-.35	1.12	89	- .49	.91	176	- .42	1.02
High-Ability	120	-.12	1.13	119	- .41	1.07	239	- .26	1.10
Low-Ability	55	-.78	.82	55	- .97	1.20	110	- .87	1.02
Total Group	176	-.33	1.09	174	- .58	1.14	349	- .46	1.11

Three-Way Analysis of Variance

Source of Variation	Sum of Squares	DF	Mean Square	*F*	p^1
Main Effects	33.84	3	11.28	9.79	.001
Ability Group	27.67	1	27.66	23.99	.001
Test	.42	1	.42	.36	.999
KR	5.63	1	5.63	4.88	.026
Two-Way Interactions	3.90	3	1.30	1.13	.340
Ability Group x Test	2.71	1	2.71	2.35	.122
Ability Group x KR	.17	1	.17	.15	.999
Test x KR	1.02	1	1.02	.89	.999
Three-Way Interaction					
Ability Group x Test x KR	.07	1	.07	.06	.999
Residual	393.15	341	1.15		
Total	430.96	348	1.24		

[1]Estimated probability of error in rejecting null hypotheses.

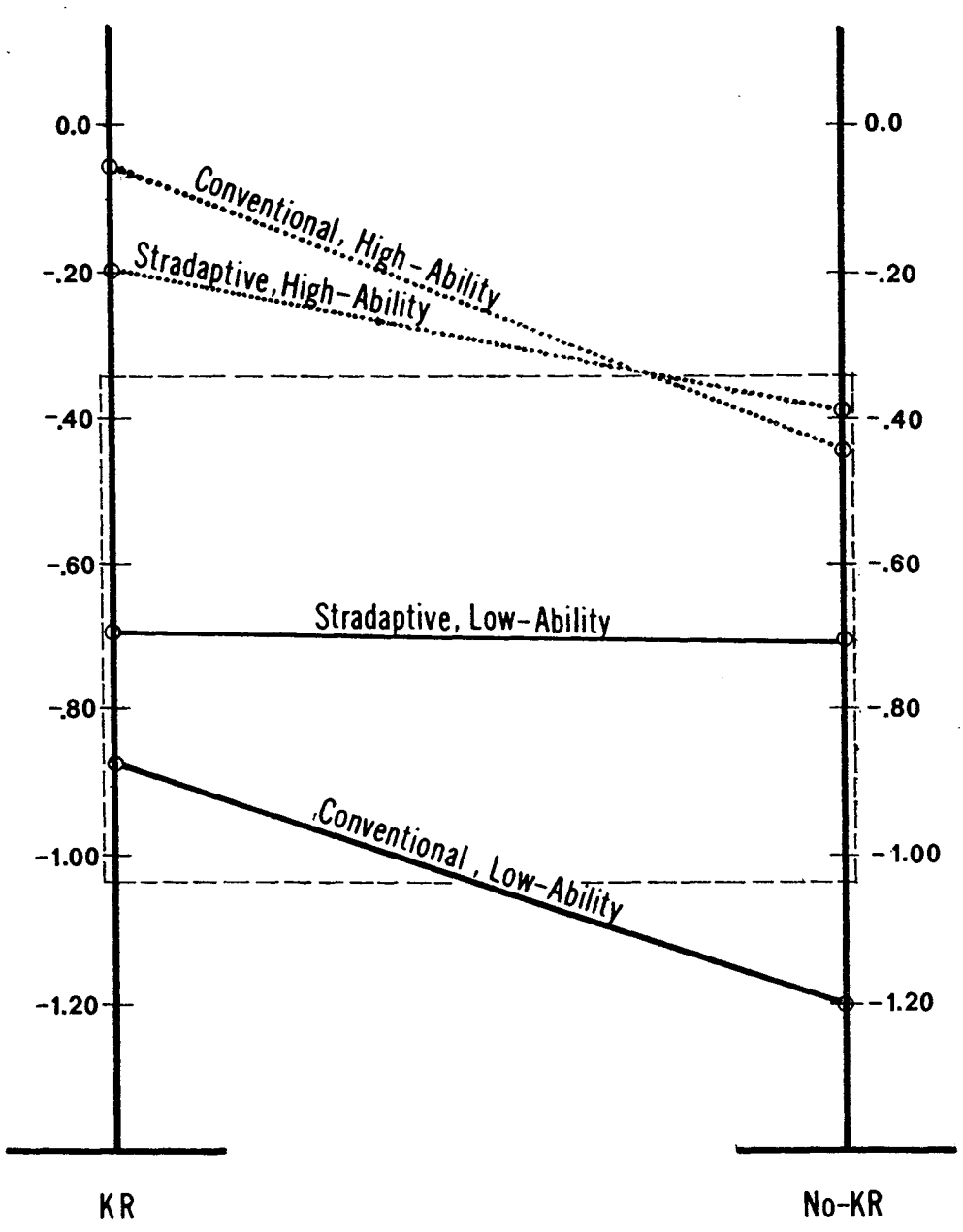
performance levels of low-ability subjects were high under KR conditions *and* on the stradaptive test in general. It seems, therefore, that the performance of the high-ability group was enhanced when KR was given regardless of testing strategy, while performance of the low-ability groups was improved under either KR conditions or by administration of an adaptive test. Interestingly, the conditions under which low-ability subjects performed most poorly, i.e., the conventional test administered without KR, are the conditions typical of most standard group-testing procedures.

Discussion

The results of this study indicate that variations in ability test administration procedures have significant effects on test scores. Specifically, significantly higher levels of performance

Figure 1

Mean Maximum Likelihood Ability Estimation as Function of Testing Strategy, KR, and Ability Group



were observed when examinees were provided with immediate knowledge of results concerning the correctness/incorrectness of each item response. Low-ability examinees obtained higher test scores when administered an adaptive rather than a conventional ability test.

Correspondent with the rationale for examining the effects of KR and adaptive testing on test scores was the finding that under conditions most nearly resembling standard practice in ability testing (i.e., conventional tests administered without KR), the performance levels of both high- and low-ability subjects were significantly lower than were performance levels observed under other testing conditions. Traditional ability testing procedures may yield scores which underestimate an individual's level on the trait of interest.

As a result of the differential effects of variations in testing conditions on the performance levels of high- versus low-ability examinees, there were some conditions under which the expected group differences in test scores were not found. Thus, situational conditions may affect not only the conclusions made about individuals on the basis of test scores, but the conclusions made about group differences in ability level. Therefore, in studying both individual and group differences in psychological variables, more attention must be paid to the possible impact of the conditions under which measurements are made on the obtained results.

References

Annett, J. *Feedback and human behavior.* Baltimore, MD: Penguin Books, 1969.

Bayroff, A. G. Feasibility of a programmed testing machine. (Research Study 64–3). Washington, DC: U.S. Army Personnel Research Office, 1964.

Beeson, R. O. Immediate knowledge of results and test performance. *Journal of Educational Research,* 1973, *66,* 224–226.

Betz, N. E. Prospects: Applications and psychological implications. In D. J. Weiss (Ed.), *Computerized adaptive trait measurement: Problems and prospects.* (Research Report 75–5). Psychometric Methods Program, Department of Psychology, University of Minnesota, Minneapolis, MN, 1975. (AD A018675)

Bierbaum, W. B. Immediate knowledge of performance on multiple-choice tests. *Journal of Programmed Instruction,* 1965, *3,* 19–23.

Bilodeau, E. A. & Bilodeau, I. McD. Motor-skills learning. In P. R. Farnsworth, O. McNemar & Q. McNemar (Eds.), *Annual Review of Psychology* (Vol. 12). Palo Alto, CA: Annual Reviews, 1961.

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, 397–479.

DeWitt, L. J. & Weiss, D. J. *A computer software system for adaptive ability measurement.* (Research Report 74–1). Psychometric Methods Program, Department of Psychology, University of Minnesota, Minneapolis, MN, 1974. (AD 773961)

Ferguson, R. L. & Hsu, T. *The application of item generators for individualizing mathematics testing and instruction.* (Report 1971/14). Pittsburgh, PA: University of Pittsburgh, Learning Research and Development Center, 1971.

Heald, H. M. *The effects of immediate knowledge of results and correlation of errors and test anxiety upon test performance.* Doctoral dissertation, University of Nebraska, 1970. (University Microfilms No. 70–17, 724).

Lord, F. M. Some test theory for tailored testing. In W. H. Holtzman (Ed.), *Computer-assisted instruction, testing, and guidance.* New York: Harper & Row, 1970.

Samejima, F. A comment on Birnbaum's three-parameter logistic model in the latent trait theory. *Psychometrika,* 1973, *38,* 221–233.

Sattler, J. M. & Theye, F. Procedural, situational, and interpersonal variables in individual intelligence testing. *Psychological Bulletin,* 1967, *68,* 347–360.

Spencer, R. E. & Barker, B. An applied test of the effectiveness of an experimental feedback answer sheet. Research Report No. 293. Urbana, IL: University of Illinois, Measurement and Research Division, 1969.

Strang, H. R., & Rust, J. O. The effects of immediate knowledge of results and task definition on multiple-choice answering. *The Journal of Experimental Education,* 1973, *42,* 77–80.

Sweet, R. C. & Ringness, T. A. Variations in the intelligence test performance of referred boys of dif-

fering racial and socioeconomic backgrounds as a function of feedback or monetary reinforcement. *Journal of School Psychology,* 1971, *9,* 399–409.

Terman, L. M. & Merrill, M. A. *Stanford-Binet Intelligence Scale manual for the third revision form L-M.* Boston, MA: Houghton Mifflin, 1960.

Wechsler, D. *Manual for the Wechsler Adult Intelligence Scale.* New York: Psychological Corporation, 1955.

Weiss D. J. *The stratified adaptive computerized ability test.* (Research Report 73-3). Psychometric Methods Program, Department of Psychology, University of Minnesota, Minneapolis, MN, 1973. (AD 768376)

Weiss, D. J. & Betz, N. E. *Ability measurement: Conventional or adaptive?* (Research Report 73-1). Psychometric Methods Program, Department of Psychology, University of Minnesota, Minneapolis, MN, 1973. (AD 757788)

Zontine, P. L., Richards, H. C. & Strang, H. R. Effect of contingent reinforcement on Peabody Picture Vocabulary Test performance. *Psychological Reports,* 1972, *31,* 615–622.

Acknowledgements

This research was supported by contracts N00014-67-A-0113-0029, NR150-343, and N00014-76-C-0243, NR150-382 from the Office of Naval Research, Personnel and Training Research Programs, David J. Weiss, Principal Investigator.

Author's Address

Nancy E. Betz, Department of Psychology, The Ohio State University, 1945 N. High Street, Columbus, OH 43210