Practice Effects with Traditional Mental Test Items

Hilda Wing
Office of Personnel Management

A nationwide Federal employment program for recent college graduates required applicants to take a multiple abilities test battery. The abilities, each assessed by a separate test part, were Verbal, Judgment, Induction, Deduction, and Number. To equate alternate forms, a sixth test part was included in the test battery. This part could be an additional, parallel version of one of the five ability test parts. At the first test administration one form (A) was used operationally, and each of the five parts of two alternate forms (B and C) was administered to a randomly selected subgroup of test takers. Small but consistent score increases from the first test form to the second were observed. The greatest effects were for Induction and Deduction, next largest for Number, and least for Verbal and Judgment. At two subsequent administrations the order of alternate form administration was reversed (B and A, C and A), providing a counterbalanced design to assess the effects of alternate form, samples, and practice. Data from 66,303 test takers supported the hypotheses of practice effects. These data suggest that practice is most effective for item types constructed according to specific rules, next effective for test parts subject to speededness, and least effective for test parts tapping general information.

The effects of practice on performance on standardized tests of cognitive ability is an old yet venerable topic. Greene's (1941) review, perhaps the most thorough discussion of practice effects, has been succeeded by texts with far less information about this topic (for example, Anastasi, 1976; Cronbach, 1970). A possible explanation for the recent neglect may be traced to the pervasiveness of standardized testing over the past 40 years. Sufficiently high levels of test sophistication have been achieved in both test construction and test taking such that practice has minimal impact (see Angoff, 1971a). For example, the Scholastic Aptitude Test (SAT) has been constructed to minimize such effects. Those who repeat the SAT from their junior to senior year in high school may expect a modest improvement in their scores varying from one-half to two-thirds of the standard error of measurement (Donlon & Angoff, 1971), that is, from .20 to .25 standard deviation units.

Score increases that do occur may be unreliable at the individual level if the test has high internal consistency and alternate forms reliability (Cronbach & Furby, 1970; Stanley, 1971). This militates against identifying correlates of score change, either potential causes such as coaching (Fremer & Chandler, 1971) or possible outcomes such as changed criterion-related validity (Donlon & Angoff, 1971). Special study of individual test repeaters who show large score changes may demonstrate, as in Jacobs (1966), that certain types of potential causes of score change are plausible, such as intensive study or

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 4, No. 2 Spring 1980 pp. 141–155

the lack of it and physical and psychological responses to testing conditions. However, only to the extent that research incorporates independent variation of potential causes will explanation replace plausibility. It may be that those individuals whose scores are deviant from expectation are more likely to search for, recall, and publicize reasons for such deviance. Further, to the extent that test repeaters are self-selected, generalizations about their score changes are limited. For example, SAT repeaters, approximately one-third of all SAT candidates, appeared to be more able than nonrepeaters (Donlon & Angoff, 1971).

Not all hypotheses about score changes are without reasonable support. Greene (1978) emphasized the importance of the test taker's motivation as well as stimulus characteristics of the test. Positive score changes could be expected in "tests where a generalized rule or method could be learned" (Greene, 1941, p. 621). Secondly, variations in score change might be reflecting error in an equating process used to obtain comparable scores for alternate forms. The reverse might also occur: Practice could introduce error into the equating process (Angoff, 1971b, pp. 574-575). However, a test data collection system designed to provide data for equating may also provide data pertinent to violations of requirements for equating. The research described here began when a formal equating system produced data pointing towards practice effects in the administration of two alternate test forms.

In 1974 the U.S. Civil Service Commission (whose employment selection functions have been assumed by its successor agency, the Office of Personnel Management) introduced the Professional and Administrative Career Examination, generally known as the PACE (McKillip, Trattner, Corts, & Wing, 1977). This examination, directed towards recent college graduates, includes a multiple abilities test battery, which is administered several times each year on a nationwide basis. The requirement for alternate forms led to the introduction of a formal scaling and equating system (Wing, 1974, 1975a, 1975b).

The written test battery was designed to have six parts. There was one operational part for each of five abilities, constant for all test takers at a given nationwide administration of the same alternate form. In addition, there was a sixth, nonoperational, unscored part which could vary during the same nationwide administration from one test taker to the next and which was used to equate alternate forms of the five ability tests or to pretest new items. One test taker would be administered only one nonoperational section at a given administration. The same alternate operational form of the battery could include from 15 to 25 different nonoperational test parts, packaged in a spiraled fashion so that assignment of a specific nonoperational form was random.

Initially, several alternate but parallel forms of the battery were developed. The first operational administration of one of these forms included all five parts of each of two other forms. That is, 10 randomly selected samples of test takers at the first nationwide administration had taken two alternate forms of the same ability test part in addition to one alternate form of each of the other four test parts. The values of the test part statistics, calculated for equating purposes, showed an unexpected pattern: Typically, the raw score mean for an alternate form taken second was higher than the mean for the form taken first. This could have occurred because of practice effects from the first form to the second or because the alternate forms were not parallel (the first form was more difficult than the second). The consistency in the size and direction of the mean discrepancies, combined with knowledge of the care taken in assembling equivalent test forms, suggested the implausibility of a hypothesis about lack of parallelism. Further data collection and analyses were required to understand and correct the equating data.

Method

Research Participants

Randomly selected samples of PACE examinees at three nationwide test administrations

provided the available data for item and test analyses. These included 42,921 tested during November 1974; 7,994 tested during May 1975; and 15,388 tested during November 1975. The total was 66,303. Other data about PACE applicants provided demographic information indicating that approximately 60% were male, over 80% were college seniors or graduates, and less than 30% were over 30 years old. In terms of their major field of study, they were probably not a representative sample of all college graduates, since there were fewer individuals majoring in the liberal arts and sciences and more individuals majoring in business and management among PACE examinees than among college students in general (Wing, 1976).

Abilities Assessed by the Written Test

The test battery assessed five abilities as described by French (1951) and by French, Ekstrom, and Price (1963). The item format was multiple-choice, five alternatives.

Verbal. Items 1 to 15 were vocabulary (synonyms), and Items 16 to 30 were reading comprehension. For the latter, each item included a different paragraph and the correct answer was a repetition, a paraphrase, or essentially a restatement of the paragraph.

Judgment. This part consisted of only one item type, comprehension. A paragraph was presented and the most plausible of the five alternatives was to be selected. The solution tapped general knowledge not included in the original paragraph.

Induction. Items 1 to 15 were letter series problems, and Items 16 to 30 were geometric classifications or analogies. The latter consisted of two sets of geometric symbols with two or more symbols in each set. The second set had a question mark replacing one symbol; the examinee was to select the most appropriate symbol to complete the analogy from one of the five alternatives.

Deduction. Items 1 to 15 were tabular completion items. Tables or charts were presented that had missing values, to be deduced from the remaining information in the table. Items 16 to 30 were inference items. A statement or paragraph was to be accepted as true, and the correct answer was to be derived from the statement without drawing upon outside information. (These items contrasted with the comprehension items measuring Judgment.)

Number. Items 1 to 5, 11 to 15, and 21 to 25 required straightforward arithmetic computation (e.g., addition, multiplication), and the remaining 15 items were word problems requiring arithmetic reasoning. The test materials were at about the eighth-grade level and had been designed to minimize the importance of verbal abilities.

Verbal processing was required by Judgment and to some extent by Deduction, as well as by Verbal itself. Judgment, and to some extent Verbal, tapped general information and knowledge. Induction was tested in a completely nonverbal way and there were low verbal requirements for Number. Induction, Deduction, and to some extent Number are also termed general reasoning abilities (Horn, 1976).

Test Materials and Procedure

The written test battery consisted of six separately timed parts, each with the same time limit of 35 minutes. The assignment of a sixth part to a test taker was random, subject to the constraint that each different nonoperational part was to be assigned (approximately) equally often.

At each test administration the first two test parts were presented in one test booklet, the second four parts in another. In this research the sixth, variable nonoperational section, was in Position 2 in the first test booklet. In Position 1 was Verbal. Judgment, Induction, Deduction, and Number were in Positions 3, 4, 5, and 6, respectively, comprising the contents of the second test booklet. That is, only for Verbal did the operational form precede the nonoperational. Test takers were to complete the test parts in order and were to work on only one test part at a time. Total testing time was 3½ hours; there was

an additional 10- to 20-minute break between the two test booklets.

Three different alternate forms were used operationally: Form A during November 1974, Form B during May 1975, and Form C during November 1975. Each of the five operational parts of Forms B and C was used as a nonoperational part during November 1974; each of the five operational sections of Form A was used as a nonoperational part during May 1975 and during November 1975. The total numbers of nonoperational parts were 15 for November 1974, 21 for May 1975, and 15 for November 1975.

Experimental Design

There were (at least) two different, but not mutually exclusive, hypotheses to explain the obtained statistics: Either the alternate form taken first was more difficult or practice on the first facilitated performance on the second. To distinguish between these two hypotheses a similar group of test takers could be administered the same two forms but in reverse order. Comparing just two forms will not, however, eliminate equivocality, as Grant (1948) and Stanley (1955) have observed, because this latin square design confounds order with sequence effects. That is, score changes in Form B when preceded by Form A might not be the same as found in Form A when preceded by Form B.

An effective way of unconfounding requires, minimally, the increase of the number of alternate forms taken by the same individuals. Such an approach did not appear feasible here as the test plan incorporated only one nonoperational section. The alternative chosen was to select two additional groups of test takers and to administer both of two alternate forms, reversing the order of administration in the second group. Then, this procedure was replicated with another two groups of test takers and another two alternate forms. To enhance comparability of these two separate investigations, one alternate form was used in both investigations; and two groups, one for each investigation, were sampled from the same population.

An outline of the experimental design is shown in Table 1. Ten randomly selected groups from the November 1974 test takers (Form A operational) were administered one additional test part from Form B or Form C. The following May 1975, Form B was used operationally. Five randomly selected groups from this population were administered one of the five ability test parts of Form A as a nonoperational section. Study 1 used Forms A and B as administered to five groups tested in November 1974 (Sample a) and five groups tested in May 1975 (Sample b). Form C was used operationally during November 1975. At this time five (out of 15) randomly selected groups were administered one of the five ability test parts of Form A as a nonoperational section. Study 2 used Forms A and C as administered to five groups tested in November 1974 (Sample c), and five groups tested in November 1975 (Sample d). For each group in each sample in each study, test data were retained only for the two alternate forms of the same ability test part.

Since some individuals took the written test twice over the three administrations, it is possible that these data include repeaters. That is, an individual might have had Number, Form B, as a nonoperational section in November 1974 and Number, Form A, as a nonoperational section in May 1975. The probability of such overlap is very small, however. Data on the incidence of test repetition suggest that an average of from three to eight people in each group were such overlapping repeaters. This is sufficiently small not to vitiate the assumption of score independence.

Test analyses. For each individual in each sample, the number of questions answered correctly, uncorrected for guessing, was recorded for both the nonoperational test part and the corresponding alternate form operational test part. In both Study 1 and Study 2 separate analyses of variance (ANOVAs) were completed for each of the five ability tests. The design was of

Table 1
Experimental Design

Alternate Test Forms Used	Operational Test parts 1, 3 4. 5, & 6	Nonoperational Test part 2
November 1974		
Sample a	Form A	Form B
group 1		Verbal
group 2		Judgment
group 3		Induction
group 4		Deduction
group 5		Number
Sample c		Form C
group 1		Verbal
group 2		Judgment
group 3		Induction
group 4		Deduction
group 5		Number
May 1975	Form B	
Sample b		Form A
group 1		Verbal
group 2		Judgment
group 3		Induction
group 4		Deduction
group 5		Number
November 1975		
Sample d	Form C	Form A
group 1		Verbal
group 2		Judgment
group 3		Induction
group 4		Deduction
group 5		Number

test forms repeated across individuals in the two samples but in different orders, with an unweighted means analysis for the unequally sized groups (Winer, 1971). For each significant two-way (test forms by samples) interaction, the proportion of variance attributable to that interaction, ω^2 (omega squared) was calculated (Dodd & Schultz, 1973; Hays, 1973). This statistic is one means of indicating the effect of practice.

A second way of indicating the size of any practice effect is given by the statistic *H* (Angoff, 1971b, pp. 574-575). This is the average difference in mean score for both test forms, for both samples, as divided by pooled estimates of the standard deviations for the test forms over both samples. This statistic was calculated for each ability test for both of the two studies, using Equation 1.

$$H = \left\{ \frac{\frac{M_{A_b} - M_{A_a}}{s_A} + \frac{M_{B_a} - M_{B_b}}{s_B}}{2} \right\}, \quad [1]$$

where

$$s_A^2 = \frac{1}{2}\left(s_{A_a}^2 + s_{A_b}^2\right), \quad [2]$$

$$s_B^2 = \frac{1}{2}\left(s_{B_a}^2 + s_{B_b}^2\right), \quad [3]$$

A, B, are form designations,
a, b, are sample designations.

Item analyses. For each test item administered to each group in each sample, three statistics were calculated: p_{tot}, the proportion of the group answering correctly; p_{att}, the proportion of those attempting the item who answered correctly; and r_{pb}, the point-biserial correlation of the item with the total test score. Two p values were calculated because in some test parts the items towards the end were attempted by fewer and fewer people. The later items were more difficult (from pretest data), so that if the test part had been entirely a test of power, those attempting to answer an item were likely to be more capable than those omitting the item. The index p_{att} would overestimate the ease, underestimate the difficulty of the item for the total group. If, however, the test part had been speeded, then some of those failing to respond to the item could have answered it correctly if they had had sufficient time. The index p_{tot} would underestimate the ease, overestimate the difficulty of the item. To the extent that analyses of p_{tot} and p_{att} lead to the same conclusions, speededness is not a factor in test performance.

Some (e.g., Donlon & Angoff, 1971) recommend that p values be transformed before analysis because of possible nonlinear relation to an underlying continuum at extreme values. Almost all of the p values in this research fell between .20 and .80, where problems of nonlinearity are minimal. Transformation was, therefore, unnecessary.

The first calculations performed, for each test part for each group in each study, were to estimate the internal consistency of the test part using the Kuder-Richardson Formula 20 (KR-20). Values of p_{tot} were used.

Second, in both Study 1 and Study 2 separate two-factor ANOVAs of both p values were completed for each of the nine item types. Test forms were repeated in the two samples, in different orders, but variance estimates were calculated across individual items rather than individual test takers. This reduced both the degrees of freedom and the generalizability of significant effects. These analyses were the same as those performed on total test scores but with far fewer repeated measures (15 for each item type but comprehension, comprising Judgment, which had 30). For each significant interaction of Samples × Test Forms, the proportion of variance attributable to that interaction was calculated (Dodd & Schultz, 1973; Hays, 1973).

Third, for each item type in each study the practice effect statistic H was calculated, as described in Equation 1. In order to do this the mean and standard deviation of the subtests had to be estimated. Using the p values and point biserials of each test item, Gulliksen (1950, pp. 365-378) has shown how to compute the mean and standard deviation of each test part. The point biserials were calculated with the total test scores, not the subtest or item type scores, and hence include the covariance between the two different item types in a test part, as well as the variance of each type. To the extent that the item types are positively correlated, using these point biserials in Equation 5 will lead to systematic overestimates of item type variance, which in turn will lead to underestimates of the practice effect as given by H in Equation 1.

$$M = \sum p_i \quad [4]$$

$$s = \sum r_{pbi} \sqrt{p_i q_i} \quad [5]$$

where

$i = 1, 2, \ldots, n$ number of items in test,
p_i = percent answering item i correctly,
q_i = percent answering item i incorrectly, or $1 - p_i$.
r_{pbi} = point biserial correlation of item i with total test score,
$\sqrt{p_i q_i}$ = standard deviation of the item.

Results

Reliability

The obtained reliability coefficients ranged from about .65 to about .85 and were acceptable for tests of cognitive abilities of 30 items each. As might be expected, the alternate forms coefficients were slightly lower than the internal consistency coefficients. Following Lord and Novick (1968), composite reliabilities for the weighted test battery used in selection could be calculated from the test part reliabilities, variances, and covariances. Such composite reliabilities would be in the low .90's (Wing, 1977).

Test Analyses

For three test parts (Induction, Deduction, Number) the mean scores of the form taken second were always higher than the mean scores of the form taken first. Consistent changes in test variance were observed only for Number, where retest variance was greater than initial variance. Since all changes in variance were small, no formal analyses were performed. For each test part in both studies the ANOVA interaction term (Samples × Test Forms) was statistically significant, indicating the presence of practice effects. The means and standard deviations of each test part are shown in Table 2.

The sample sizes used here made it easy to obtain statistical significance. Of more value are the statistics indicating the size of the practice effect, ω^2 and H, displayed in Table 3. Practice appeared to be a minor source of variance in these data, at the most accounting for 2.5% of the total variance in Induction scores in Study 2. However, average test performance decidedly improved. For Induction, Deduction, and Number, the simple addition of practice had consistently increased scores on the average of a point or two across groups that were heterogeneous in ability.

The standard errors of measurement for each test part were estimated from the reliability coefficients. The effects of practice for Induction, Deduction, and Number Varied from about one-half to three-quarters of the size of the corresponding standard error. Over the five parts of the total test battery, the data suggest an average increase of almost five raw score points after practice on all five parts, if the effects are uncorrelated. The standard error of measurement for the sum of the five test scores (from Wing, 1977) is approximately 6.5 raw score points. That is, the predicted practice effect for the total test was approximately three-quarters of the standard error of measurement, a figure slightly greater than that observed for SAT repeaters. The variability of the change scores here would be more than the variability due to measurement error, a finding again similar to that observed with SAT repeaters (Donlon & Angoff, 1971).

Item Statistics

ANOVAs of item difficulty values generally had nonsignificant interaction terms, thus failing to demonstrate effects of practice on the different item types. This lack of support is probably attributable to the reduced power of the item ANOVAs based on the low values for degrees of freedom. The proportion of accountable

Table 2
Summary Statistics of Total Test Scores

Samples	Test Part				
	Verbal	Judgment	Induction	Deduction	Number
	Study 1: Forms A and B				
Sample a (November 1974)					
Sample size	4,694	3,957	4,018	4,096	4,165
Test taken first					
Mean	17.61	19.97	16.26	16.61	11.87
Standard deviation	6.10	4.22	5.21	4.94	4.40
Test taken second					
Mean	17.65	21.03	17.43	17.39	14.04
Standard deviation	5.52	4.42	5.21	5.62	4.66
Sample b (May 1975)					
Sample size	1,637	1,626	1,608	1,584	1,539
Test taken first					
Mean	17.03	20.49	15.32	16.36	13.01
Standard deviation	5.42	4.70	5.10	5.85	4.49
Test taken second					
Mean	17.37	19.90	17.84	18.03	13.44
Standard deviation	6.15	4.22	5.44	4.91	4.83
	Study 2: Forms A and C				
Sample c (November 1974)					
Sample size	4,721	4,220	4,333	4,350	4,367
Test taken first					
Mean	17.61	20.16	16.53	16.96	12.61
Standard deviation	6.14	4.28	5.54	5.57	4.47
Test taken second					
Mean	17.31	20.98	17.39	17.61	14.03
Standard deviation	5.96	4.43	5.36	5.64	4.77
Sample d (November 1975)					
Sample size	3,138	3,142	3,091	3,055	2,962
Test taken first					
Mean	16.35	20.47	16.34	16.31	12.74
Standard deviation	6.01	4.64	5.51	5.99	4.53
Test taken second					
Mean	17.12	20.28	19.35	18.90	13.75
Standard deviation	6.22	4.15	5.89	5.32	4.84

Note. For Verbal, the operational form was administered before the nonoperational form. For all other test parts the operational form was administered after the nonoperational form.

Table 3
Practice Effect Statistics of Total Test Scores

Practice Effect Statistic	Test Part				
	Verbal	Judgment	Induction	Deduction	Number
ω^2: Practice effect as percent of variance					
Study 1	0.02	0.05	2.26	0.93	1.40
Study 2	0.03	0.11	2.55	1.74	1.43
H: Practice effect in standard deviation units					
Study 1	0.04	0.05	0.35	0.23	0.28
Study 2	0.04	0.07	0.34	0.29	0.26

variance for those item types where the interaction term was significant, as well as the practice effect statistic H for each item type, are in Table 4. These latter estimates are intriguing even though they are probably underestimates of the "true" practice effect. Note that the statistics based on p_{att} are, with the exception of reading comprehension, smaller than those based on p_{tot}. That is, eliminating omitted responses in the attempt to eliminate the effect of speededness reduced the effects of practice. (The exception may reflect little more than rounding error.)

To illustrate the effects of speededness, scatterplots of the two kinds of p values may be constructed. For each kind, the proportion of the group responding correctly to an item is plotted depending on whether that item appeared in the alternate form taken first or taken second. To the extent that these item points deviate from the 45° line, practice effects are present. To the extent that such deviations are different for item points of different item types, practice affects these item types differently. To the extent that the scatterplots for p_{tot} differ from those for p_{att}, practice affects also the usage of time in tests that are speeded.

Two representative scatterplots are shown in Figures 1 and 2. In Figure 1 are displayed the scatterplots for Deduction, Form A, Study 2. The November 1974 group (Sample c) encountered this form after experience with the alternate Form C, whereas the November 1975 group (Sample d) took Form A prior to taking Form C. The left scatterplot is of the proportions based on entire groups, and the right scatterplot is of proportions based on only those who attempted to answer the items. In Figure 2 are displayed the two scatterplots for Form B of Number in Study 1, in which the November 1974 group (Sample a) encountered Form B before Form A and the May 1975 group (Sample b) took Form B after Form A.

Practice Effects

The effects of practice varied from one test part to the next, reflecting changes in the skill of dealing with different item types as well as in the speed of answering test questions.

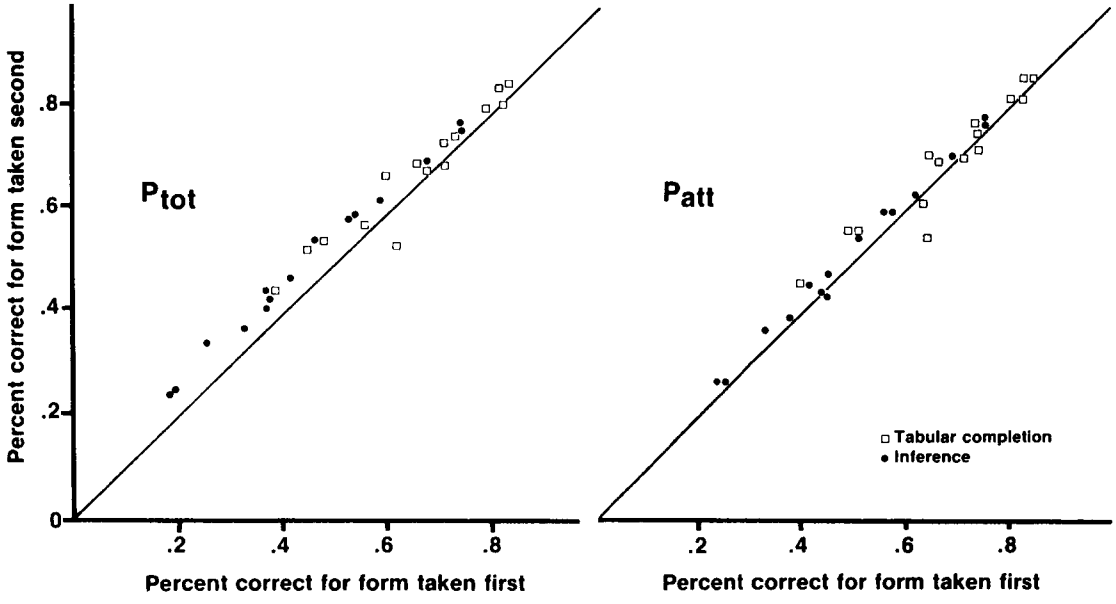
Verbal. Practice improved scores slightly, more for the easier reading comprehension items than the more difficult and earlier vocabulary items. The reading comprehension items were less likely to tap general information than were the vocabulary items. There appeared to be little effect of speededness.

Judgment. Practice led to slight improvement. Elimination of omitted responses led to no change in summary statistics, indicating that speededness was not a factor.

Figure 1

Proportion Answering Item Correctly, for Each Item, on Deduction, Form A for Those Taking Form A First (Sample d, November 1975) Versus Those Taking Form A Second (Sample c, November 1974)

On the left are the proportions for the total samples (p_{tot}); on the right are the proportions for those not omitting the item (p_{att}). The 45° diagonal or identity relation is drawn to facilitate comparison.



Induction. The first and easier item type, letter series, was greatly affected by practice. Scores for the second item type, geometric classifications, were also improved with practice. The elimination of omitted responses reduced the size of the practice effect statistics by about one-third, suggesting that practice improved both accuracy and speed of response.

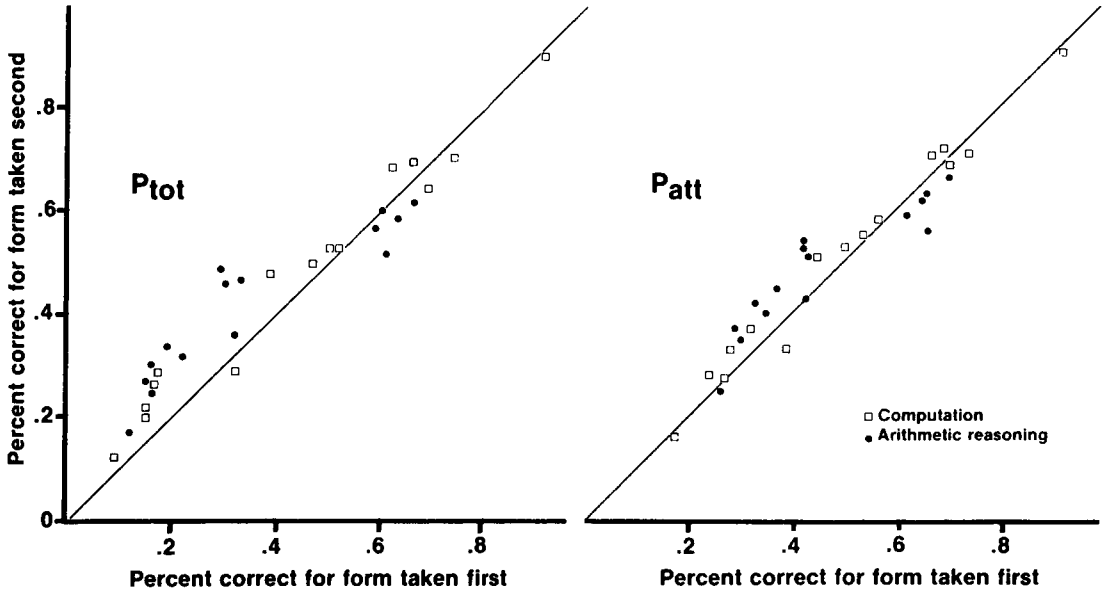
Deduction. The more difficult later inference items appeared to be more subject to practice effects than the easier and earlier tabular completion items. Speededness may also be a factor, as illustrated in Figure 1. Eliminating the omitted responses reduced the practice effect more for the inference items, suggesting that part of the improvement with practice may be attributable to reaching more of these items.

Number. Table 4 shows sizable practice effects for both computation and arithmetic reasoning items, effects which were markedly reduced when omitted responses are eliminated. The form of the speededness here is different from that in Deduction, however, as Figure 2 shows. In the form taken second, more of the more difficult arithmetic reasoning items were answered correctly, whereas fewer of the easier items, either computation or arithmetic reasoning were correct. However, when those who omitted responding were eliminated, as shown in the right of Figure 2, this differential effect was removed and *all* items appeared equally subject to a slight effect of practice. That is, the effects of practice on these two item types appeared to be more on the speed or pacing of re-

Figure 2

Proportion Answering Item Correctly, for Each Item, on Number, Form B for Those Taking Form B First (Sample a, November 1974) Versus Those Taking Form B Second (Sample b, May 1975)

On the left are the proportions for the total samples (p_{tot}); on the right are the proportions for those not omitting the item (p_{att}). The 45° diagonal or identity relation is drawn to facilitate comparison.



sponse rather than on increasing skill in answering such items correctly.

Discussion

The research reported here was initiated as a response to aberrant operational test equating data. The evidence of two counterbalanced studies with two alternate forms was more supportive of the practice effects hypothesis than of the hypothesis of differential difficulty of forms. The effects appeared to be larger for those item types (letter series, geometric classifications, arithmetic reasoning, tabular completion, inference, reading comprehension) that are solvable by systematic application of general problem-solving skills than for those item types (vocabulary, comprehension) solvable by application of general previously acquired information. Practice may also have led to better use of available time in tests where speededness may have been important, such as observed here in that test part evaluating the Number ability. Further, such effects were present in random samples of all test takers, not just in self-selected samples of test repeaters.

That practice effects are frequent can be documented by an elderly, scattered, yet extensive literature. As noted above, Greene's (1941) review remains the most thorough and his hypothesis of the importance of generalized procedures in item solution in predicting practice effects remains viable. More recent evidence shows practice effects in verbal analogies (Colver

Table 4
Practice Effect Statistics of Estimated Item Type Scores

Practice Effect Statistic		Item Type[a] Vocabulary	Reading Comprehension	Letter Series	Geometric Class.	Tabular Completion	Inference	Computation	Arithmetic Reasoning
ω^2: Practice effect as percent of variance[b]									
p_{tot}:	Study 1	--	--	8.58	--	--	1.50	--	1.88
	Study 2	--	--	6.68	--	--	2.12	--	1.88
p_{att}:	Study 1	--	--	6.24	--	--	--	--	--
	Study 2	--	--	5.11	--	--	--	--	--
H: Practice effect in estimated standard deviation units									
p_{tot}:	Study 1	-0.01	0.09	0.39	0.30	0.20	0.29	0.21	0.34
	Study 2	0.01	0.08	0.39	0.29	0.25	0.35	0.16	0.34
p_{att}:	Study 1	-0.01	0.10	0.30	0.21	0.14	0.06	0.07	0.10
	Study 2	0.01	0.09	0.32	0.22	0.17	0.12	0.00	0.11

[a]Since the test part Judgment was constructed from only one item type (comprehension), the total test statistics as presented in Table 3 are appropriate.
[b]Calculated only when Samples times Test Forms interaction term reached statistical significance at or beyond $p = .01$.

& Spielberger, 1961; Spielberger, 1959), reading comprehension (Vernon, 1962), letter series (Anastasi & Drake, 1954), syllogisms (Johnson-Laird & Steedman, 1978), and various tests of spatial ability (Blade & Watson, 1955; Krumboltz & Cristal, 1960). Score changes have been observed on more heterogeneous measures such as the Wechsler Intelligence Scale for Children (Quereshi, 1968), the Wechsler Adult Intelligence Scale (Elwood, 1972), the General Aptitude Test Battery (Droege, 1966), and even the SAT (Frankel, 1960; French & Dear, 1959; Whitla, 1962). Practice effects appear to be specific to the particular test or test items practiced (Gagne & Paradise, 1961; Nevo, 1976; Woodrow, 1946). The data reported here indicate that practice effects are not limited to those who choose to retake a test. It is possible, of course, that practice effects for self-selected repeaters would be different from those for randomly selected samples.

More troublesome are the qualifications presented by specific characteristics of this research. The first is the "experimental" nature of the nonoperational section. To the extent that test takers knew that the section did not "count" and, consequently, that they did not give it the same attention and effort that they gave to the operational version, the effects of practice observed here are not accurate. Scores on the nonoperational form would have been systematic underestimates of ability. A second consideration is that of test speededness. The influence of practice in speeded tests has been observed elsewhere. Rate of response has been isolated as a distinct factor of test score variance (Lord, 1956; Myers, 1952; Sternberg, 1977). Mollenkopf (1950; 1960) found that additional test time did

not improve scores in a test of verbal analogies but did lead to improvement in a test of arithmetic reasoning. Some type of speed-accuracy trade-off was possible for the latter test but not the former.

The scattered nature of the practice effects literature emphasizes the difficulty of study and understanding. Such change scores are typically small (Angoff, 1977) as well as individually unreliable (Cronbach & Furby, 1970), suggesting the necessity of large samples to provide adequate statistical power. Stating that practice changes test scores is not providing an explanation, let alone predicting which kinds of tests for which examinees produce which kinds of practice effects. The latter kind of information is most important for operational testing programs in which both test takers and test users wish to know the predicted outcomes of test repetition. This information could serve to discount, or uphold, advertisements of testing "cram schools" by including documented correlates of score change.

Of equal if not more importance are more detailed, perhaps experimental, studies to untangle what, in practice, is being changed. Specifically, the item types that appear to be most subject to practice effects are those used to assess fluid as opposed to crystallized intelligence (Horn, 1966; 1976). The vulnerability of tests of this generalized trait to "short-period fluctuations" has been noted (Horn, 1966, p. 559). Could it be a characteristic of fluid intelligence to show such practice effects? How is differential susceptibility to practice effects, of item types and of test takers, related to different aspects of intelligence?

That such questions will be difficult to answer does not minimize their importance, particularly since there is much other research demonstrating differences in profiles of cognitive ability scores among groups differing along the socially relevant dimensions of sex, race, and ethnicity (Wing, 1979). For example, some groups may demonstrate lower levels of performance on tests of fluid intelligence as compared to tests of crystallized intelligence. The theory behind such distinctions is weak (see, for example, Mandler & Stein, 1977; Sherman, 1978) and the evidence is inconsistent (Arvey, 1972). If the item types and tests used were also differentially subject to practice effects, then the inconsistency in results becomes explicable. Dubin, Osburn, and Winick (1969) found that extra practice on certain speeded tests was most helpful to lower socioeconomic white and higher socioeconomic black male teenagers. Replication and extension of such investigations are obviously necessary.

References

Anastasi, A. *Psychological testing* (4th ed.). New York: Macmillan, 1976.

Anastasi, A., & Drake, J. D. An empirical comparison of certain techniques for estimating the reliability of speeded tests. *Educational and Psychological Measurement,* 1954, *14,* 529-540.

Angoff, W. H. (Ed.) *The College Board admissions testing program: A technical report on research and development activities relating to the Scholastic Aptitude Test and Achievement Test.* New York: College Entrance Examination Board, 1971. (a)

Angoff, W. H. Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, DC: American Council on Education, 1971. (b)

Angoff, W. H. Personal communication, June 1977.

Arvey, R. D. Some comments on culture fair tests. *Personnel Psychology,* 1972, *25,* 433-448.

Blade, M. F., & Watson, W. S. Increase in spatial visualization scores during engineering study. *Psychological Monographs,* 1955, *69* (Whole No. 397).

Colver, R., & Spielberger, C. D. Further evidence of a practice effect in the Miller Analogies Test. *Journal of Applied Psychology,* 1961, *45,* 126-127.

Cronbach, L. J. *Essentials of psychological testing* (3rd ed.). New York: Harper & Row, 1970.

Cronbach, L. J., & Furby, L. How we should measure "change"—or should we? *Psychological Bulletin,* 1970, *74,* 68-80.

Dodd, D. H., & Schultz, R. F., Jr. Computational procedures for estimating magnitude of effect for some analysis of variance designs. *Psychological Bulletin,* 1973, *79,* 391-395.

Donlon, T. F., & Angoff, W. H. The Scholastic Aptitude Test. In W. H. Angoff (Ed.), *The College Board admissions testing program: A technical report on research and development activities relating to the Scholastic Aptitude Test and Achieve-*

ment Test. New York: College Entrance Examination Board, 1971.

Droege, R. C. Effects of practice on aptitude scores. *Journal of Applied Psychology,* 1966, *50,* 306-310.

Dubin, J. A., Osburn, H., & Winick, D. M. Speed and practice: Effects on Negro and White test performance. *Journal of Applied Psychology,* 1969, *53,* 19-23.

Elwood, D. L. Automated versus face-to-face intelligence testing: Comparison of test-retest reliabilities. *International Journal of Man-Machine Studies,* 1972, *4,* 363-369.

Fremer, J., & Chandler, M. O. Special studies. In W. H. Angoff (Ed.), *The College Board admissions testing program: A technical report on research and development activities relating to the Scholastic Aptitude Test and Achievement Test.* New York: College Entrance Examination Board, 1971.

Frankel, E. Effects of growth, practice, and coaching on scholastic aptitude test scores. *Personnel and Guidance Journal,* 1960, *38,* 713-719.

French, J. W. The description of aptitude and achievement tests in terms of rotated factors. *Psychometric Monographs* 1951, (No. 5).

French, J. W., & Dear, R. E. Effect of coaching on an aptitude test. *Educational and Psychological Measurement,* 1959, *19,* 319-330.

French, J. W., Ekstrom, R. B., & Price, L. A. *Manual for kit of reference tests for cognitive factors.* Princeton, NJ: Educational Testing Service, 1963.

Gagne, R. M., & Paradise, N. E. Abilities and learning set in knowledge acquisition. *Psychological Monographs,* 1961, *75* (Whole No. 518).

Grant, D. A. The latin square principle in the design and analysis of psychological experiments. *Psychological Bulletin,* 1948, *45,* 427-442.

Greene, E. B. *Measurement of human behavior.* New York: The Odyssey Press, 1941.

Greene, E. B. Personal communication, July 1978.

Gulliksen, H. *Theory of mental tests.* New York: Wiley, 1950.

Hays, W. L. *Statistics for the social sciences* (2nd ed.). New York: Holt, Rinehart, & Winston, 1973.

Horn, J. L. Integration of structural and developmental concepts in the theory of fluid and crystallized intelligence. In R. B. Cattell (Ed.), *Handbook of multivariate experimental psychology.* Chicago: Rand McNally, 1966.

Horn, J. L. Human abilities: A review of research and theory in the early 1970s. *Annual Review of Psychology,* 1976, *27,* 437-485.

Jacobs, P. I. Large score changes on the Scholastic Aptitude Test. *Personnel and Guidance Journal,* 1966, *45,* 150-156.

Johnson-Laird, P. N., & Steedman, M. The psychology of syllogisms. *Cognitive Psychology,* 1978, *10,* 64-99.

Krumboltz, J. D., & Cristal, R. E. Short-term practice effects in tests of spatial aptitude. *Personnel and Guidance Journal,* 1960, *38,* 385-391.

Lord, F. M. A study of speed factors in tests and academic grades. *Psychometrika,* 1956, *21,* 31-50.

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

Mandler, J. M., & Stein, N. L. The myth of perceptual defect: Sources and evidence. *Psychological Bulletin,* 1977, *84,* 173-192.

McKillip, R. H., Trattner, M. H., Corts, D. B., & Wing, H. *The Professional and Administrative Career Examination: Research and development* (PRR-77-1). Washington, DC: U.S. Civil Service Commission, 1977. (NTIS No. PB 268 780)

Mollenkopf, W. G. An experimental study of the effects on item-analysis data of changing item placement and test time limit. *Psychometrika,* 1950, *15,* 291-315.

Mollenkopf, W. G. Time limits and the behavior of test takers. *Educational and Psychological Measurement,* 1960, *20,* 223-230.

Myers, C. T. The factorial composition and validity of differently speeded tests. *Psychometrika,* 1952, *17,* 347-352.

Nevo, B. The effects of general practice, specific practice, and item familiarization on change in aptitude test scores. *Measurement and Evaluation in Guidance,* 1976, *9,* 16-20.

Quereshi, M. Y. Practice effects on the WISC subtest scores and IQ estimates. *Journal of Clinical Psychology,* 1968, *24,* 79-85.

Sherman, J. A. *Sex-related cognitive differences.* Springfield, IL: Thomas, 1978.

Spielberger, C. D. Evidence of a practice effect on the Miller Analogies Test. *Journal of Applied Psychology,* 1959, *43,* 259-263.

Stanley, J. C. Statistical analysis of scores from counterbalanced tests. *Journal of Experimental Education,* 1955, *23,* 187-207.

Stanley, J. C. Reliability. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, DC: American Council on Education, 1971.

Sternberg, R. J. *Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities.* Hillsdale, NJ: Erlbaum, 1977.

Vernon, P. E. The determinants of reading comprehension. *Educational and Psychological Measurement,* 1962, *22,* 269-286.

Whitla, D. K. Effect of tutoring on scholastic aptitude test scores. *Personnel and Guidance Journal,* 1962, *41,* 32–37.

Winer, B. J. *Statistical principles in experimental design* (2nd. ed.). New York: McGraw-Hill, 1971.

Wing, H. *The scoring system of the Professional and Administrative Career Examination (PACE)* (TM-74-2). Washington, DC: U.S. Civil Service Commission, 1974.

Wing, H. *Scaling and equating series of Test 500 FY 75* (Operations paper). Washington, DC: U.S. Civil Service Commission, 1975. (a)

Wing, H. *Scaling and transmutation procedures, Test 500* (Operations paper). Washington, DC: U.S. Civil Service Commission, 1975. (b)

Wing, H. *Normative data for the Professional and Administrative Career Examination (PACE): FY 1975* (TM-76-15). Washington, DC: U.S. Civil Service Commission, 1976. (NTIS No. PB 268 786).

Wing, H. *Stability characteristics of alternate forms of a test battery* (TM-77-7). Washington, DC: U.S. Civil Service Commission, 1977. (NTIS No. PB 280 958).

Wing, H. *Profiles of cognitive ability of different racial/ethnic and sex groups on a multiple-abilities test battery* (TM-79-7). Washington, DC: U.S. Office of Personnel Management, 1979. (*Journal of Applied Psychology,* in press).

Woodrow, H. The ability to learn. *Psychological Bulletin,* 1946, *53,* 147–158.

Acknowledgments

Portions of this research were presented at the 19th annual meeting of The Psychonomic Society, San Antonio, TX, November 1978. Helpful discussions were held with E. Elizabeth Stewart, William H. Angoff, and Yosef H. Pavlov. Comments by W. H. Angoff, L. S. Buck, R. J. Karren, and J. D. Kraft on earlier versions of this report were very helpful. The opinions expressed are those of the author and do not necessarily reflect those of the Office of Personnel Management.

Author's Address

Send requests for reprints or further information to Hilda Wing, Personnel Research and Development Center, Room 3226, Office of Personnel Management, Washington, DC 20415.