Retesting in Selection: A Meta-Analysis of Practice Effects for Tests of Cognitive Ability

John P. Hausknecht

Cornell University

Jane A. Halpert

Nicole T. Di Paolo

Meghan O. Moriarty Gerrard

DePaul University

In Press: *Journal of Applied Psychology*

Author Note:

John P. Hausknecht, Department of Human Resource Studies, Cornell University; Jane A. Halpert, Nicole T. Di Paolo, and Meghan O. Moriarty Gerrard, Department of Psychology, DePaul University.

We thank Bridgette Harder and Christopher Mason for assistance with data coding. We are also grateful to Jose Cortina and two anonymous reviewers for their helpful comments on previous versions of this manuscript.

Correspondence concerning this article should be addressed to John P. Hausknecht, School of Industrial and Labor Relations, Department of Human Resource Studies, 388 Ives Hall, Cornell University, Ithaca, New York, 14853. E-mail: jph42@cornell.edu.

Abstract

Previous studies indicate that as many as 25-50% of applicants in organizational and educational settings are retested with measures of cognitive ability. Researchers have shown that practice effects are found across measurement occasions such that scores improve when these applicants retest. This study uses meta-analysis to summarize the results of 50 studies of practice effects for tests of cognitive ability. Results from 107 samples and 134,436 participants revealed an adjusted overall effect size of .26. Moderator analyses indicated that effects were larger when practice was accompanied by test coaching, and when identical forms were used. Additional research is needed to understand the impact of retesting on the validity inferences drawn from test scores.

KEYWORDS: personnel selection, cognitive ability testing, practice effects, meta-analysis

Retesting in Selection: A Meta-Analysis of Practice Effects for Tests of Cognitive Ability

Over 40 years ago, researchers asked whether retesting applicants for necessary skills and abilities was justified in personnel selection (van der Ries, 1963). Scholars in educational settings had been studying similar questions concerning the consequences of retesting since the early 1920s (e.g., Richardson & Robinson, 1921). More recently, research has begun to provide empirical answers to questions with regard to the prevalence, outcomes, and implications of retesting practices and policies for those tested in organizational and educational contexts. Over 20 years have passed since this growing body of literature was reviewed empirically (Kulik, Kulik, & Bangert, 1984). Advancements in test development, testing technology, and the experience of test takers suggest that an updated assessment of this literature may yield different findings than earlier reviews, and may identify moderator variables that help explain why effects differ across settings or populations. Therefore, the purpose of this study is to synthesize existing research on practice effects for tests of cognitive ability using meta-analysis.

Retesting Prevalence, Policies, and Recommendations

The available evidence suggests that a sizable proportion of test takers do indeed return for an additional test. For example, approximately 50% of high school students retake the SAT I (Nathan & Camara, 1998), roughly 1/3 of MCAT candidates are repeat test takers (Koenig & Leger, 1997), about 40% of test takers have completed the Swedish SAT more than once (Cliffordson, 2004), and nearly 40% of candidates for admission to medical school in Belgium were repeating the test for at least the second time (Lievens, Buyse, & Sackett, 2005). In organizational settings, available data suggest that retesting may be less common. For example, 11% of retail managers who were candidates for promotion chose to retest (Tuzinski, Laczo, &

Sackett, 2005) and approximately 25% of law enforcement candidates appeared for a second administration of a cognitive ability test (Sin, Farr, Murphy, & Hausknecht, 2004).

Given the prevalence of repeated assessment, test developers and policymakers in government and professional organizations have issued a wide range of rules and recommendations concerning retesting, some of which are conflicting. For example, the website for the company that delivers the Miller Analogies Test (MAT) notes that "if an examinee's second (or most recent) test score is 25 points or greater than the first (or most recent previous) score, the second score is invalidated" (Harcourt Assessment, 2005). Carretta, Zelenski, and Ree (2000) reported that the United States Air Force prohibits retesting altogether on the Basic Attributes Test, which is part of a test battery used to select military pilots. Finally, professional guidelines for employee selection state that "employers should provide opportunities for reassessment and reconsidering candidates whenever technically and administratively feasible" (Society for Industrial and Organizational Psychology, 2003), and "every test taker should have a fair chance to demonstrate his or her best performance on an assessment procedure…consider retesting or using alternative assessment procedures before screening the individual" (U.S. Department of Labor, 1999, Section 4). Accordingly, many of the major standardized test publishers (e.g., SAT, GRE, LSAT, GMAT) allow retesting, although they often establish minimum time periods for the test-retest interval.

Why Candidates Choose to Retest

A variety of factors underlie candidates' decisions to retest. In academic settings, individuals may retest during the application process in order to enhance their candidacy for admission to the college or graduate program of their choice (Brounstein & Holahan, 1987). In other cases, candidates may retest after failing to gain admission to graduate or professional

programs in their desired field (Cliffordson, 2004; Lievens et al., 2005). In addition, retesting may be prompted by academic program requirements stipulating that candidates achieve a given score on a standardized test (Lane, Penn, & Fischer, 1966). In employment contexts, candidates likely choose to retest because their initial score did not lead to a job offer (Hausknecht & Howard, 2004; Thomas, Busciglio, & Goldenberg, 2004) or because they believed that retesting would enhance their chances for a promotion (Tuzinski et al., 2005). Thus, in academic and employment settings, retesting is prompted by some level of candidate discontent with either the initial test score or the outcome of the application process. Finally, in research settings, candidate retesting is often prompted by the researcher, which suggests that the decision to retest does not reflect the same degree of volition or will on the part of the candidate as it might in operational selection contexts (Messick & Jungeblut, 1981).

From a theoretical standpoint, expectancy-based models of motivation help explain test-taking motivation in general (Sanchez, Truxillo, & Bauer, 2000), and may also explain the decision to retest in operational selection settings. This approach is predicated on the notion that motivation is a function of the valence, instrumentality, and expectancy of the situation (Vroom, 1964). In particular, valence refers to the degree of attractiveness or desirability of a particular outcome. In the testing context, candidates who view the employment opportunity or academic program as positively valent would be motivated to retest because of the satisfaction that is expected to come from admission to that particular program, institution, or organization. Instrumentality refers to the belief that performance will lead to a preferred outcome, and in the case of selection retesting, implies that candidates will retest because they believe that earning a higher score will result in an offer of employment or admission. Finally, expectancy refers to an individual's subjective assessment of the probability of attaining a particular outcome. In

selection settings, candidates will choose to retest because they believe that they have the capacity to earn higher scores upon retesting as a result of their effort. In summary, theory and research suggest that the decision to retest may be motivated by valence, instrumentality, and expectancy beliefs, which compel candidates to retest in an effort to enhance the likelihood that they will be offered an opportunity to join the organization or academic program of their choice.

Effect of Retesting on Subsequent Test Performance

Researchers have repeatedly shown that performance on tests of cognitive ability improves across measurement occasions. Score gains have been found in organizational (e.g., Hausknecht, Trevor, & Farr, 2002), educational (e.g., Powers & Rock, 1999), clinical (e.g., Basso, Carona, Lowery, & Axelrod, 2002), and research settings (e.g., Woehlke & Wilder, 1963). These practice effects, defined here as changes in a person's test score from one administration to the next, have been found with some regularity, yet the magnitude of the effect differs across studies and may be due to variability in the number of participants included in each study and variability in the reliability of the measure used in the research. Therefore, the first goal of this study is to synthesize existing research using meta-analysis in order to provide the best estimate of the magnitude of practice effects for measures of cognitive ability after correcting for variation that is due to sampling and measurement error artifacts.

Hypothesis 1: Scores on tests of cognitive ability will increase across administrations.

Explanations for Practice Effects

Given the support for the prevalence of practice effects on measures of cognitive ability, it is important to understand *why* practice effects occur across administrations. Developing a better theoretical understanding of the reasons for practice effects could help explain variability in the magnitude of score gains, and has implications for how retesting should be managed in

major testing programs. Explanations for practice effects include reduced anxiety (Messick & Jungeblut, 1981), memory of previous responses (Kulik, Kulik, et al., 1984), actual development of abilities (Anastasi, 1981), enhanced test taking strategies (Sackett, Burris, & Ryan, 1989), and regression to the mean (Campbell & Kenny, 1999). One or more of these explanations is related to the five moderators of practice effects that are assessed in this study, including test coaching, formal instruction, study context, test form, and cognitive ability dimension[1]. In the sections that follow, we present theory and hypotheses that establish the rationale for testing these effects. In addition, we address regression to the mean as a possible alternative explanation for score gains, and discuss how retesting could alter the validity inferences that may be drawn about test scores.

Mere Repetition

Before discussing potential moderators, it is important to note that practice effects may be attributable to *mere repetition*, meaning that scores on subsequent administrations could change as a result of having completed a cognitive ability test in the past. In other words, scores may improve from one administration to the next absent any type of intervention such as test coaching or formal instruction. Perhaps these gains are attributable to familiarity with the testing environment or enhanced understanding of item types on the part of candidates, but in any event, the mere repetition scenario serves as a baseline by which to evaluate the relative efficacy of other interventions (e.g., coaching, instruction) toward improving test scores. Note that mere repetition refers to the act of completing additional cognitive ability tests that are officially proctored and scored by an independent organization, researcher, or test administrator. When candidates take additional "practice tests" on their own, these activities are more consistent with test coaching activities (described below).

[1] Several additional moderators were considered in this study, such as gender, age, and test name, although this information was missing from many of the primary studies. The results of these analyses are available from the first author.

Test Coaching

All studies of practice effects necessarily involve two measurements, with some level of time lapse between test administrations, which suggests that the magnitude of score gains may be related to differences in the intervening activities that occur between tests. Test coaching is one type of intervention that has been shown to influence score gains in some cases (Anastasi, 1981; Messick & Jungeblut, 1981). Test coaching involves instruction aimed at improving test scores, and can be considered along a continuum, as described by Messick and Jungeblut (1981):

> What has come to be called *coaching* is here considered to fall anywhere in the broad range between these two extremes of practice and instruction, entailing some combination of test familiarization, drill-and-practice with feedback, training in strategies for specific item formats and for general test taking (including advice on pacing, guessing, and managing test anxiety), subject matter review, and skill-development exercises. (p. 192)

Thus, this line of reasoning suggests that, to the extent that coaching programs emphasize test-taking strategies over subject matter review and skill building, scores improve over time because test coaching is efficacious in developing the test-taking skills of candidates, which allows them to perform better on a subsequent administration. Note that because cognitive ability tests generally are not designed with the purpose of assessing test-taking strategies, the validity of the measure would be compromised to the extent that the content of the test coaching program emphasizes test-taking strategies versus skills that are assessed by the test. When candidates improve test scores without improving these underlying skills, any inferences that are made about subsequent test scores will be less valid.

Messick and Jungeblut (1981) operationally defined test coaching as the amount of student contact time spent in coaching and found that coaching time was logarithmically related to score gains. Their results showed that for the SAT-Verbal and SAT-Math, the amount of coaching contact time required to produce score gains greater than 20 to 30 points (on an 800-point scale) rivaled that of full-time schooling. In other words, test coaching appeared to provide some initial benefit to candidates, yet also yielded diminishing returns such that greater and greater amounts of time in coaching were required to produce the same effect. Their findings lend support to the notion that coaching may be efficacious, but also reveal an important boundary condition of the effect that suggests very large gains would be difficult to obtain.

Hypothesis 2: Practice effects will be positively related to the amount of candidate contact time in test coaching.

Formal Instruction

The degree to which candidates are actively participating in formal instruction is another likely influence on the magnitude of practice effects. In their research on the development of cognitive ability, Cahan and Cohen (1989) contend that their findings "unambiguously point to schooling as the major factor underlying the increase of intelligence test scores" (p. 1239). Particularly when studies involve students who are enrolled in relevant coursework, score gains may reflect real increases in knowledge, skills, and abilities that are important for the test rather than, or perhaps in addition to, effects due to coaching (Messick & Jungeblut, 1981). Formal instruction is defined here as long-term, intensive study, such as schooling, that is targeted at developing widely applicable intellectual skills, work habits, and problem-solving strategies (Anastasi, 1981). Whereas cognitive ability was once thought to be fixed and stable throughout the lifespan, empirical studies have shown that improvements in the construct are possible,

although such enhancements generally require significant investments on the part of the individual (Anastasi, 1981). Formal instruction differs from test coaching in that the purpose of the activities is to develop general skills or knowledge as opposed to "teaching to the test." Because instruction is actually improving the abilities that are relevant to the test, score gains may be attributable to true changes in the construct of interest, which would render validity inferences unchanged regardless of whether initial or retest scores are considered.

Hypothesis 3: Practice effects will be positively related to the amount of candidate contact time in formal instruction.

Study Context

In studies of practice effects conducted in operational contexts, candidates choose to retest of their own volition, whereas in research settings, the experimenter is responsible for creating the retesting environment. This critical difference in study context introduces self-selection factors that may influence the magnitude of practice effects that are observed because of differences in candidate motivation (Arvey, Strickland, Drauden, & Martin, 1990). For example, candidates who are completing tests as part of the admission process to college or graduate school, and job applicants who are competing for a position within an organization, may be motivated to enroll in formal coaching programs or to engage in targeted skill development because there is a highly desired outcome at stake. On the other hand, participants who are completing multiple tests for research purposes, whether within an organization, school, or laboratory setting, may be less motivated to enhance their scores upon retesting because they have little to gain or lose in these contexts. Moreover, in operational contexts, those who repeat the test have an incentive to repeat the test *and* to perform well during the second administration because the distal outcome to be gained is relatively long-term in nature and is largely dependent

upon test scores. In contrast, participants in research settings may be motivated to participate because of course credit, but this type of reward is relatively short-lived and often will be delivered regardless of test performance.

In sum, candidates in operational contexts differ from participants in research studies in several meaningful ways. The lasting consequences of testing in operational contexts, coupled with the fact that mere participation is insufficient to earn the desired outcome, suggests that practice effects in operational settings will be greater than those found in research contexts.

Hypothesis 4: Practice effects will be related to study context such that the magnitude of score gains will be greater in operational contexts than in research contexts.

Test Form

Practice effects may also vary depending on whether the second test is the exact same test as the first, or whether it is an alternate form. One of the primary reasons that test developers create alternate forms of a test is to combat memory effects that could unduly influence test scores across administrations. Cook and Campbell (1979) suggested that "familiarity with a test can sometimes enhance performance because items and error responses are more likely to be remembered at a later testing session" (p. 52). In a repeat testing context, such benefits will accrue only to candidates who receive the exact same test, which suggests that practice effects will be larger when identical forms are used rather than alternate forms.

The relationship between test form and practice effects is complicated by the fact that the time interval between tests will vary across studies[2]. Clearly, a time lapse of several years does not allow memory effects to influence practice effects in the same way that would be expected from intervals that are relatively brief. To the extent that memory effects are responsible for practice effects, robust evidence from cognitive psychology indicates that these benefits will

[2] We thank an anonymous reviewer for suggesting this possibility.

dissipate over time because of memory decay (Best, 1999), which suggests that for relatively long intervals, identical forms may be more similar to alternate forms. This pattern would be revealed empirically for identical tests as a negative association between time interval and practice effects.

Hypothesis 5a: Practice effects will be related to test form such that the magnitude of score gains will be greater for identical forms than for alternate forms.

Hypothesis 5b: Practice effects for identical forms will be negatively related to the time interval between tests.

Cognitive Ability Dimension

There is a long history of research related to understanding the structure of cognitive abilities (Vernon, 1961). Although a full review of this literature is beyond the scope of this study, there is evidence to support a hierarchical structure that consists of verbal, quantitative, and analytical sub-facets of general cognitive ability or "*g*" (Carroll, 1993; Roth, Bevier, Bobko, Switzer, & Tyler, 2001). Tests of verbal ability are designed to assess language skills, and often include analogies, vocabulary, or reading comprehension items. Tests of quantitative ability measure mathematics skills, and typically include items that involve arithmetic computation, mathematical reasoning, or other numerical problem solving. Tests of analytical ability assess the ability to use logic, and may include items that measure inductive and deductive reasoning (Wing, 1980).

Practice effects may differ among cognitive ability dimensions because candidates can systematically apply general problem solving skills to some items and not others. For instance, quantitative and analytical test items and some verbal test items (e.g., word analogies) can be solved by applying general problem solving skills. However, the majority of verbal test items

(e.g., sentence completion) require the acquisition of new information (e.g., learning new vocabulary words). To the extent that repeat administrations engender familiarity with procedural features of the testing environment and perhaps reduce anxiety (Anastasi, 1981), candidates may be able to work through quantitative and analytical problems more quickly, which should yield higher scores during the second administration. These benefits would not apply to most measures of verbal ability, however, because these items require that candidates actually learn new information between tests. Wing (1980) studied practice effects for these three dimensions in a study of recent college graduates, and because tests were given consecutively on the same day, candidates did not have an opportunity to acquire new information. She found that practice effects were much larger for item types that could be solved by applying specific rules (i.e., quantitative measures) than for those tapping general information and knowledge (i.e., verbal measures). Empirical data from subsequent research also reveal greater gains for tests of quantitative ability when compared with tests of verbal ability (Brounstein & Holahan, 1987; Powers & Rock, 1999).

> *Hypothesis 6:* Practice effects will be related to cognitive ability dimension such that score gains will be larger for tests of quantitative and analytical ability when compared to measures of verbal ability.

Regression to the Mean

Another plausible explanation for practice effects is regression to the mean, which occurs when there is an imperfect correlation between two variables and when extreme groups are selected for study. The effect is such that extreme scores on one variable (e.g., initial test scores), whether high or low, tend to be paired with scores that are less extreme on the other variable

(e.g., repeat test scores). The result is that high scores will regress downward toward the mean, whereas low scores will regress upward toward the mean.

Regression to the mean is important for this study because those who decide to retest may constitute an extreme group. In particular, candidates who score high enough to gain admission or secure a job offer will not appear for retesting, thereby creating a group of candidates who scored, on average, below the mean of the entire group that tested initially. Given this natural form of extreme group selection, and the often imperfect correlation between scores across administrations, regression to the mean will occur such that scores of candidates who repeat the test would be expected to regress upward toward the mean upon retesting, thereby providing an alternative explanation for practice effects. This situation may be especially likely when candidates choose to retest because illness or some other transient factor hindered their initial performance. Very few studies of practice effects report adequate data to empirically test regression to the mean effects. However, we were able to examine the results of two studies that did report the necessary information.

Effects of Retesting on Validity

Finally, given that the purpose of testing often is to forecast performance in some other domain, it is important to understand how retesting alters the validity inferences that may be drawn from test scores when candidates generate more than one score on a measure of cognitive ability. Although the effects of retesting on subsequent test performance are well-researched, very little empirical evidence examines how retesting may alter validity inferences concerning important criteria such as job performance or success in graduate school. In fact, we could locate only three studies that indirectly address whether and how the validity inferences may be altered

by repeat testing. Given the relative scarcity of studies on the topic, the findings from these studies are reviewed below in narrative format rather than via meta-analysis.

First, Allalouf and Ben-Shakhar (1998) related scores on the Israeli Psychometric Entrance Test to scores on a matriculation exam that contained open-ended questions about different topical areas that were required as part of the high school curriculum such as English, Math, and Biblical studies. The authors found that the correlation between entrance test scores and matriculation exam scores did not differ depending on whether the entrance test score was from the second or third administration (data from the first test administration were not reported). The authors noted that universities typically used the matriculation scores in conjunction with scores on the cognitive measure. Thus, the two scores might actually represent relationships among predictors rather than between predictors and criteria. Further, the use of matriculation exam scores is somewhat uncharacteristic of the criteria used in most validation studies where job performance or some other future-oriented measure of success is sought.

Second, Lievens et al. (2005) examined validity coefficients relating cognitive ability test scores to GPA to determine whether the first or second test score was a better predictor of GPA. Results revealed no significant differences between the two coefficients. However, when examining scores on a science knowledge test, they did find that the second test score was more predictive of GPA than the first. The authors cited the use of GPA as the criterion as one limitation of their study and suggested that, “future research should examine whether our results generalize to employment settings with job performance as the criterion” (p.1004).

Third, Hausknecht et al. (2002) examined relationships between repeated measures of cognitive ability and training performance for 1,515 candidates that were eventually selected into positions in law enforcement. The authors reported a validity coefficient of .31 when the entry-

gaining test score was correlated with training performance and a value of .27 when the initial test score was used. Although these results provide some support for the idea that there is little impact of retesting on validity inferences, the fact that the dependent variable was training performance, rather than job performance, leaves the issue unresolved. In fact, the conclusion reached by the authors was that "though the posthire training performance finding is suggestive, how repeat testing relates to posttraining performance on the job remains unknown" (p. 253).

In summary, the results of these three studies suggest that there may be little or no difference in validity coefficients depending on which test score is used. However, test users may be interested in criteria that are different from those examined in these few studies. For example, no study to date has determined how retesting might influence validity inferences when the criterion is job performance, which is arguably the very reason that organizations choose to test candidates using measures of cognitive ability. The effect of retesting on validity is also likely to be dependent on the underlying factors that cause practice effects to occur. When score gains reflect improvements in test taking abilities without parallel growth in the underlying abilities measured by the test, inferences drawn regarding test scores from subsequent administrations will be compromised. However, inferences will remain stable to the extent that candidates develop relevant skills between tests or overcome test-irrelevant factors that limit performance, such as test anxiety. Finally, it is important to note that the discussion of retesting and validity is complicated by the fact that there may be inter-individual differences in the rate of practice or learning among candidates, or in the degree to which certain candidates can surmount test anxiety between administrations. Regardless of how the overall validity coefficient may or may not differ in these circumstances, inferences about any one individual may change from one test to the next.

Method

Literature Search

To locate relevant journal articles, an extensive search of the *PsycINFO* (1900 – December 2004), ABI/INFORM (1970 – December 2004), and Academic Search Elite (1984 – December 2004) computerized databases was performed using variations and combinations of the following keywords: *retesting, repeat testing, practice effects,* and *test coaching*. Reference lists from the most recent meta-analytic investigation of practice effects and test coaching on tests of cognitive ability (i.e., Kulik, Bangert-Drowns, & Kulik, 1984; Kulik, Kulik, et al., 1984) were scanned to locate studies. The computerized *Social Science Citation Index* database was used to identify articles that referenced either Kulik et al. meta-analysis. Finally, manual and computerized searches of the Society for Industrial and Organizational Psychology (1999 – 2005) and the Academy of Management (1999 – 2004) academic conference programs revealed several additional papers and presentations that were applicable to this study. These search methods yielded approximately 200 articles, which were reviewed by the authors to determine whether they could be included in the analysis using the criteria described below.

Criteria for Inclusion

The criteria for inclusion in the meta-analysis were: (1) a measure of cognitive ability was collected from the same sample on at least two occasions; (2) the available data were sufficient to calculate an effect size; and (3) the study participants were from a normal (i.e., not cognitively impaired or hospitalized) population. Because the intention was to assess the effects of practice on measures of cognitive ability, knowledge tests of specific content domains were excluded from the meta-analysis (e.g., biology, psychology), as were reaction time studies in the cognitive psychology literature. These decision rules also exclude “posttest only” designs (e.g.,

Powers, 1985) because all studies had to include scores from two or more test administrations. The final data set included 107 independent samples extracted from 50 articles.

Meta-Analytic Procedure

The authors and two doctoral students in industrial and organizational psychology coded the set of articles following a half-day training session designed to familiarize raters with the coding process, provide opportunities for discussion and interpretation of the coding sheet, and allow practice ratings of sample articles. During training, raters were given the coding sheet and a process guide that provided definitions of study characteristics (e.g., alternate form vs. identical form) and instructions for coding individual studies. After reviewing these materials as a group, each rater independently coded two sample articles. Upon completion, each item on the coding sheet was reviewed, discussed, and any differences were resolved. Following the training session, the coding sheet and process manual were revised based on the group discussion, and redistributed along with the articles that were retained for inclusion in the study.

For the actual coding of articles, a pair of raters independently read and coded each article using the revised coding sheet. Many of the variables that were coded did not require judgment, but were nonetheless dual-coded to ensure accuracy of transcription. Two of the variables analyzed as moderators did require judgment calls, study context and test form. Agreement was assessed using Cohen's kappa, and values of .88 for study context and .75 for test form were found. Each pair of raters then subsequently met to discuss any differences and to reach consensus on all decision points.

Effect size calculations. In order to calculate the overall effect size, which represents the average gain in cognitive ability test scores from one administration to the next, the formula for Cohen's *d* was used:

$$d = (M_{\text{Time 2}} - M_{\text{Time 1}})/SD_{\text{pooled}}$$

This value reflects the difference in group means from Time 1 to Time 2, divided by the pooled standard deviation, which can be approximated by taking the square root of the sample-weighted average of the within-group variances, where the respective N-1 values form the weights and N-2 is used as the denominator. A reviewer recommended that this formula be used instead of the formula for correlated designs (cf. Smither, London, & Reilly, 2005) because the latter formula includes the pre-post correlation (Dunlap, Cortina, Vaslow, & Burke, 1996), which leads to confounding the effect size with the test under study, and precludes accurate conclusions about the magnitude of score gains that can be expected.

Only one effect size per sample was calculated. When necessary, multiple dimensions (e.g., verbal, quantitative) were averaged to produce a single, overall effect size. When results were reported for multiple samples within the same paper, and the samples were statistically independent from one another, they were not averaged prior to the meta-analysis (Hunter & Schmidt, 2004). Unless stated by the authors to the contrary (e.g., Allalouf & Ben-Shakhar, 1998), it was assumed, and in 12 studies explicitly stated by the authors, that the data reported for the first test administration did in fact represent the candidate's first test.

Each effect size was corrected for sampling error and measurement error using the formulas outlined by Hunter and Schmidt (2004). For samples that did not report a reliability estimate ($k = 67$), a statistic of .82 was imputed, which reflects the average of those values that were reported in the remaining samples ($k = 40$).

Moderator analyses. The guidelines outlined by Cortina (2003) for the detection and estimation of moderators were followed. Hypotheses were formulated *a priori* and strategies for estimation were selected based on the nature of the variables. Both subgroup analysis and

weighted least squares (WLS) regression were used to test hypothesized relationships between moderator variables and effect size estimates (Steel & Kammeyer-Mueller, 2002).

All studies involved mere repetition because of the decision rule to include only studies involving two or more administrations of a cognitive ability test. In terms of coding information for the WLS regression moderator analyses, *coaching time* was coded as the total number of hours of candidate contact time with coaching, *instruction time* was coded as the total number of days of candidate contact time with formal instruction, *test form* was coded as to whether the second test was an identical or alternate form (identical forms coded higher), and *study context* was coded to reflect whether the study was conducted in an operational (i.e., academic/employment) or research setting (operational coded higher). For the subgroup analysis, test coaching was coded as to whether or not it was delivered to participants between administrations (coded "yes"/"no"), and test form and study context were coded in the same manner described above. Formal instruction was not assessed via subgroup analysis.

Results

Description of Primary Studies

Participant characteristics. The literature search for primary studies resulted in 107 usable samples (N = 134,436) drawn from 50 studies that met all criteria for inclusion in the meta-analysis. The median number of participants per sample was 91.00 (M = 1,256.41, SD = 3,647.71, Min = 10, Max = 33,969). The mean reported age of each sample was located and averaged across all samples (M = 19.32 years, SD = 11.22, Min = 5.00, Max = 56.50). The current education level that was most representative of each sample at the time of the study was coded into one of three categories: k-8th grade (28% of all samples included in the meta-analysis), high school (26%), or college (46%).

Moderators. In terms of study context, more samples were drawn from research contexts (k = 88) than from operational settings (k = 19). Less than half of the samples (k = 38) completed an identical form for both test measurements, while the majority completed an alternate form (k = 64); information was unavailable for the remaining samples (k = 5). All 107 samples had data available for two measurement points, 16 samples reported data from the second to third test administration, and 15 samples included data from the first to the third test administration.

Considering the interventions reported between tests, 23 of the 107 samples received at least some test coaching, and the median amount of candidate contact time in these 23 samples was 2.5 hours. For the remaining 84 samples, there was sufficient information contained in the primary study to conclude that test coaching was not delivered to participants in 75 samples. For the remaining 9 samples, there was insufficient information in the study to determine whether or not coaching was delivered, and these samples were not included in the test coaching results.

In terms of formal instruction, 53 of the 107 samples received at least some formal instruction between test administrations, and the median amount of candidate contact time in these 53 samples was 60 days. For the remaining 54 samples, there was sufficient information contained in the primary study to conclude that formal instruction was not delivered to participants in 49 samples, and for the remaining 5 samples, there was insufficient information in the study to make such a determination.

Other test and study characteristics. The median time interval between the first and second test administrations was 20 days (M = 134.52 days, SD = 304.67 days, *Min* = 15 minutes, *Max* = 5.18 years). Studies spanned several decades (1921-2005), with 84% published between 1960 and 2005 and 20% published between 2001 and 2005. Finally, approximately 90% of the cognitive ability tests were administered in a group setting via paper-and-pencil.

Overall Practice Effects

Hypothesis 1 predicted that scores would increase across administrations of a cognitive ability test. Results based on 107 samples revealed a sample-weighted mean *d* of .24 (see Table 1). When these values were further corrected for sampling error, the overall mean (δ) was .26. Put differently, these findings reveal that scores improved about one-quarter of a standard deviation across the first and second administrations of a cognitive ability test. Additional analyses revealed a sample-weighted mean *d* of .18 for practice effects from the second to third administration (δ = .20, k = 16), and a sample-weighted mean *d* of .51 (δ = .56, k = 15) for non-consecutive tests (i.e., Test 1 to Test 3). Hypothesis 1 was supported.

Moderators

Lipsey (2003) noted that moderator variables may be related to each other and to effect size estimates in meta-analytic research, which creates confounding that can make the interpretation of the results of any one moderator variable misleading. For example, it may appear that practice effects are higher when identical forms are used, but if test form is also associated with other design or sample characteristics, the conclusions drawn regarding test form may be spurious.

In order to address this possibility, we examined the interrelationships between four moderators included in this study (i.e., coaching contact time, instruction contact time, study context, and test form). As shown in Table 2, results indicate that the relationship between instruction time and study context was statistically significant (r = .37, p < .01), as was the relationship between coaching time and test form (r = .20, p < .05). In other words, exposure to formal instruction was greater in operational contexts and exposure to coaching was greater when identical forms were used.

Given this overlap, and the variability in sample size across studies, we entered the four moderator variables into a WLS regression equation to determine the relative contributions of each variable in predicting *d.* Each effect size was weighted via the inverse of the sampling error variance, using the mean effect size in the formulas outlined by Hunter and Schmidt (2004). In addition, given the severe skew of the coaching contact time and formal instruction time variables, a logarithmic transformation (using "started logs", where appropriate) was applied to these variables before entering them into the equation (Cohen, Cohen, West, & Aiken, 2003). These results are reported in Table 3. When all four moderators are considered simultaneously as predictors of *d*, test form and coaching time emerge as statistically significant, whereas study context and formal instruction time do not. These findings are discussed in the context of the subgroup analysis results described below.

Test coaching. Hypothesis 2 stated that the magnitude of practice effects would be positively related to the amount of candidate contact time with coaching. The results of the WLS regression displayed in Table 3 show that the coefficient for coaching contact time was statistically significant ($\beta = .26$, $p < .05$). When compared using subgroup analysis (Table 4), studies that included some form of test coaching had an overall effect size ($\delta = .70$) that was larger than the average practice effect in studies that did not include any type of test coaching ($\delta = .24$). The 95% confidence intervals for these estimates did not overlap. Hypothesis 2 was supported.

Formal instruction. Hypothesis 3 stated that practice effects would be positively related to the candidate contact time with formal schooling. The WLS regression results in Table 3 show that the coefficient for formal instruction was not statistically significant. Given the lack of

significant findings, and the continuous nature of the formal instruction variable, subgroup analysis was not performed. Hypothesis 3 was not supported.

Study context. Hypothesis 4 stated that practice effects would be larger in operational selection contexts than in research settings. The coefficient for study context shown in Table 3 did not reach conventional levels of statistical significance in the WLS regression analyses. As shown in Table 4, when assessed using the subgroup analysis method, results revealed an overall effect size for studies conducted in research contexts (δ = .24) that was similar to that found for studies conducted in operational contexts (δ = .29). Results also show that the 95% confidence interval overlapped for these estimates. Hypothesis 4 was not supported.

Test form. Hypothesis 5a stated that practice effects would be larger for identical forms than for alternate forms. The WLS regression results shown in Table 3 indicate that practice effects were greater when studies involved identical forms than when they used alternate forms (β = .37, p < .01). As shown in Table 4, when assessed using subgroup analysis, results revealed an overall effect size for identical tests (δ = .46) that was larger than that found for alternate forms (δ = .24). The 95% confidence intervals for these estimates did not overlap. Hypothesis 5a was supported.

Hypothesis 5b predicted that memory decay might cause the magnitude of practice effects to decline over time, but only for identical measures. Using WLS regression, we regressed the effect size onto the log-transformed time interval separately for identical and alternate forms. For identical measures, the coefficient relating time interval to effect size was significant and in the expected direction (β = -.51, p < .01), indicating that larger practice effects were associated with shorter time intervals. For alternate forms, the coefficient was not statistically significant (β = .22, p = .09). Hypothesis 5b was supported.

Using the raw values for time interval, we then used WLS regression to predict the time interval that would be required for the effect size for identical forms to equal that found for alternate forms. The resulting predicted value, 685 days, suggests that the effect size for identical forms, while higher than that for alternate forms, decreases over time such that practice effects for identical forms will be similar in magnitude to those found for alternate forms after an interval of just under two years.

Cognitive ability dimension. Hypothesis 6 stated that practice effects for tests of quantitative ability and analytical ability would be larger when compared to tests of verbal ability. Practice effects were largest for tests of analytical ability (δ = .32), followed by quantitative ability (δ = .30), and verbal ability (δ = .19). Although these results are in the expected direction, the overlapping confidence intervals indicate that these estimates are not statistically different from one another. Hypothesis 6 was not supported.

Regression to the mean. Bobko (2001, p. 167) provides a formula that can be used to estimate the degree of regression to the mean, $Y = r(Z_X)$, where Y equals the predicted Time 2 score in standard deviation units after accounting for regression to the mean, r equals the correlation between measurements at Time 1 and Time 2, and Z_X equals the standardized difference between the mean of the selected group at Time 1 and the mean of the total group at Time 1. Bobko gives the example that if one selects a group of underperformers who are 2 standard deviations below the mean at Time 1 (Z_X = -2.00), and the correlation between performance at Time 1 and Time 2 (r) is .80, then the predicted future score for these individuals (Y) equals .80(-2.00) or -1.60. In other words, there is an expected gain of .40 standard deviation units that is directly attributable to regression to the mean.

In this study, we located two papers that contained the data necessary to assess how much of the score gain may be attributable to regression to the mean. The formula requires knowledge of the *M* and *SD* for all candidates who took the test at Time 1, including those who chose to retest *and* those who did not. Two papers in our database provided such information (Lievens et al., 2005; Tuzinski et al., 2005).

In the Lievens et al. paper, we calculated Z_X to be -.27, meaning that the group of candidates who decided to retest scored approximately one quarter of a standard deviation below the mean at Time 1. Given the reported value of the test-retest correlation (.84), we solved the equation above to determine the predicted score for the retest group at Time 2, after accounting for regression to the mean. The result, -.23, suggests that a gain of .04 standard deviation units is expected because of regression to the mean. The actual value of *d* reported in Lievens et al. is .42, which means that .38 is not attributable to regression to the mean.

Applying this same approach to the data reported in Tuzinski et al. (2005) reveals that the retest group scored -.20, or one-fifth of a standard deviation below the mean at Time 1. The results of the formula cited above suggested that a gain of .04 standard deviation units is expected upon retesting because of regression to the mean. The actual value of *d* reported in Tuzinski et al. is .41, which means that .37 is not attributable to regression to the mean. In sum, regression to the mean appears to be partially responsible for practice effects, although the portion of the total practice effect that is attributable to this phenomenon may be fairly small (i.e., less than 10% in these cases).

Note that when extreme group selection does not occur, and the same group is tested at Time 1 and Time 2, the formula above shows there to be no regression to the mean (i.e., $Y = Z_X$). In this study, 82% of the samples included in the meta-analysis database were conducted in

research contexts where all participants were retested at Time 2, yet practice effects were still found in these settings. Thus, when practice effects are found in the absence of extreme group selection, as in the case of these studies, they are likely to be attributable to factors other than regression to the mean.

Discussion

Research indicates that as many as 25-50% of candidates who are tested in operational contexts have completed at least one previous test administration. Expectancy-based models of motivation suggest that the decision to retest is driven by candidates' dissatisfaction with their initial score and the belief that retesting will improve their chances of receiving an offer to join a desired organization or academic program. Previous research has shown that candidate retesting produces practice effects for tests of cognitive ability, and the results of this meta-analysis confirm that this finding is robust. A number of variables were expected to moderate the magnitude of these effects, and we were able to test many of these empirically in this study.

Summary and Implications

The results of the overall meta-analysis show that test scores increase approximately one-quarter of a standard deviation when assessed from the first to the second administration, and increase approximately one-fifth of a standard deviation from the second to the third administration. Finally, when assessing gains from the first to the third test administration, results reveal a mean practice effect that is slightly larger than one-half of a standard deviation. In practical terms, when based on normative data from the first test, these results suggest that a candidate who scored at the 50th percentile on the first test could be expected to score at the 60th percentile on the second test, and at the 71st percentile on the third test.

The meta-analysis results also suggest that practice effects are heterogeneous, which indicates the likely presence of moderators. We hypothesized that: (a) time spent with coaching would be positively related to practice effects, (b) time spent receiving formal instruction would be positively related to practice effects, (c) practice effects would be greater in operational contexts than in research contexts, (d) practice effects would be greater when tests were identical rather than alternate forms, and (e) practice effects would be smaller for tests of verbal ability than for measures of quantitative and analytical ability. Overall, support was mixed for these predictions. The relationship between practice effects and the amount of formal instruction received by candidates between tests was not significant, and practice effects did not differ by study context. When practice effects were analyzed separately for verbal, quantitative, and analytical dimensions of cognitive ability, score gains were found in all cases, but the analysis revealed no significant differences in the magnitude of these effects when compared across dimensions.

On the other hand, we did find support for the test coaching hypothesis, which showed that practice effects were larger when coaching was delivered between tests. While this finding appears to support the efficacy of test coaching, results also show that mere repetition accounts for some portion of practice effects, and this finding must not be overlooked when interpreting the test coaching results. Further, results from the regression analysis confirm Messick and Jungeblut's (1981) finding that practice effects are logarithmically related to candidate contact time in coaching. In other words, test coaching effects deteriorate over time such that greater and greater amounts of coaching are required to produce the same gain in test scores.

Practice effects also differed as expected by the test form used in the retesting situation. The average effect for studies using identical forms across administrations was larger than that

found for studies using alternate forms. Memory effects may account for this difference, as the time interval between tests was negatively related to the magnitude of score gains only for identical forms. The fact that practice effects still occur when alternate forms are used suggests that score gains are not simply indicative of recall effects. We also found that the observed effect size for identical forms is expected to be comparable to that for alternate forms after a time lag of nearly two years. One implication of these findings is that organizations may be able to minimize practice effects due to memory by using a minimum retest interval of at least one year.

The fact that practice effects did not differ by study context is somewhat surprising given the differences in experience, motivation, attitudes, and skills of participants across settings. Research based on student samples has been criticized for its lack of generalizability to applied ("real-world") settings, and there is some empirical evidence that results may vary between student and non-student samples (Gordon, Slade, & Schmitt, 1986). Although practice effects in each context appear to be similar in terms of the magnitude of the observed effect size, this finding does not necessarily imply that an identical patterning of determinants generates practice effects. Additional research is necessary to study whether there are substantive differences in the conditions that cause practice effects in research and operational contexts.

Regression to the Mean

For the two studies that reported complete data, we estimated that less than 10% of the total effect size could be attributed to regression to the mean. We want to emphasize the importance of considering the selection ratio in any study of practice effects because, to the extent that this group of repeat test takers deviates from the overall mean, effects attributable to regression to the mean would be larger. Note that the selection ratio in the Lievens et al. study, approximately 30%, is somewhat higher than what is typically found in education and

employment settings. For example, one study reported that the average selection ratio across 253 doctoral programs in psychology was 11% (Chernyshenko & Ones, 1999), while selection ratio estimates in organizational contexts have been estimated at 10 to 30% (Scullen, Bergey, & Aiman-Smith, 2005). The selection ratio is positively related to the magnitude of regression to the mean effects in the sense that high selection ratios produce a relatively extreme group of candidates who are available to retest (i.e., only those who did poorly at Time 1), which makes regression to the mean more likely. On the other hand, low selection ratios produce a retest candidate pool whose average initial score is nearly identical to that for the entire group that completed the first test, which should minimize regression to the mean. In addition, as the reliability of the cognitive ability measure approaches unity, the effects of regression to the mean are also lessened. Lastly, the influence of regression to the mean on practice effects may be more serious in operational selection contexts because candidates typically return for retesting only when they perform poorly on the first test (thereby creating a relatively extreme group), whereas most studies conducted in research settings retest all candidates, making this form of self-selection less likely.

In any event, researchers who study practice effects should consider regression to the mean as a plausible explanation for score gains. Given that regression to the mean involves extreme group selection, a related concern for researchers should be to explore reasons for poor performance during the initial administration. To the extent that candidates cite transient factors such as illness or fatigue as limiting their opportunity to perform during the initial test, performance should be enhanced upon repeated testing for some individuals, although it is important to note that regression to the mean effects will still occur whenever extreme groups are selected, regardless of the reasons underlying poor performance. Very few of the studies in this

database included any direct assessment of state-related factors that could predict the decision to retest.

Future Research

Although meta-analysis research can answer many questions, such a methodology also reveals gaps in the literature that are worthy of future study. One pressing area that could benefit from careful empirical research is to examine whether and how retesting changes the validity inferences that can be drawn from test scores. When academic institutions or organizations use tests for the purpose of admission or selection, their primary interest is in the validity of the test scores. That is, how well does the test predict later performance on the job, in graduate school, or in college? The question of whether these organizations should be using initial test scores or repeat test scores as a predictor of applicants' future performance has rarely been studied. As described above, the various explanations for practice effects have different implications for the validity inferences that may be drawn regarding test scores. In the case of test coaching that is aimed at improving test-taking skills, the candidate's first test score may be the better predictor of success because any gains in test performance reflect construct-irrelevant improvements that do not extend into the criterion domain. In the case of instruction that improves the underlying ability, the most recent test score may be the best indicator of success, given that performance should reflect the actual standing on the construct of interest. Of the 107 samples included in this meta-analysis, none of the studies included job performance data so that validity information could be compared across tests. In addition, very few studies reported data for different subgroups (e.g., ethnic background). Studies that help answer questions about test fairness and adverse impact in retesting situations would be especially useful in addition to those that address validity concerns.

Another possible avenue for future research is to continue evaluating the notion that certain item types may be more or less susceptible to practice effects. For example, Powers (1986) found that score gains were positively related to the length and complexity of the test's instructions, and were negatively related to the response time allotted per question. Item types that had the same set of response options across items, and items that had four response options rather than five were also associated with larger effects. Two additional variables, the average difficulty of test items and whether or not sample items were provided to candidates, were not significantly related to the effect size. Powers concluded that modifications to instructions, time limits, and item formats may render tests less susceptible to practice effects. Overall, very little research has examined practice effects at the item level.

The evidence provided from validity research in repeat testing situations can also inform decision makers about how to handle the score gains that are typically found across administrations. Anastasi and Urbina (1997) suggested that "allowance for such gains should be made when interpreting test scores" (p. 25), but were not specific about the nature and meaning of such allowances. In practice, policies vary widely as to how scores based on retesting will be considered in admission or selection settings. Some organizations may take the best of the two scores, while others may take the average or most recent score. With such little guidance from the empirical literature, these policies and practices are capricious at best.

Finally, little research, if any, has explored applicants' reactions to retesting. Most of the procedural justice rules outlined by Gilliland (1993) are relevant to retesting situations. For example, the opportunity to perform rule suggests that applicants will perceive the testing process as being more fair and valid when they believe they have had a chance to demonstrate relevant skills and abilities during the process. In addition, the opportunity for reconsideration

rule is relevant to retesting situations because applicants may perceive the process more favorably when they know that they have the option to retest if their initial attempt is unsuccessful. The variability in retesting policies and recommendations outlined earlier is sure to produce very different reactions from candidates in operational settings. This also represents a rich area for future research.

Limitations

Although the authors of a number of primary studies were careful to describe how they verified that the test sequence was accurate, it is possible that the data reported for some candidates included in this study for the first test may actually reflect a second or third administration. For example, it is common for candidates to prepare for standardized admission tests by completing practice tests independently before appearing for the first "official" test administration. In organizational settings, it is also possible that candidates may have taken pre-employment cognitive ability tests for another organization prior to completing the measures that were labeled the first and second test. To the extent that candidates were engaged in such preparation or practice before participating in the studies reported here, the magnitude of the practice effects reported in this study may be underestimates of the actual effect.

Conclusion

Candidate retesting is fairly common, and this study shows that practice effects are also perhaps just as widespread. An average overall effect size of .26 was estimated based on results from 107 samples. Moderator analyses show that effects were larger when candidates received test coaching and when identical forms were used across administrations. Remaining research needs include studying the effects of retesting on validity, which has implications for the meaning of practice effects, and assessing how repeated testing influences applicants' perceptions of test fairness and validity.

References

* References marked with an asterisk indicate studies included in the meta-analysis.

*Allalouf, A., & Ben-Shakhar, G. (1998). The effect of coaching on the predictive validity of scholastic aptitude tests. *Journal of Educational Measurement, 35,* 31-47.

Anastasi, A. (1981). Coaching, test sophistication, and developed abilities. *American Psychologist, 36,* 1086-1093.

Anastasi, A., & Urbina, S. (1997). *Psychological testing.* Upper Saddle River, NJ: Prentice-Hall.

Arvey, R.D., Strickland, W., Drauden, G., & Martin, C. (1990). Motivational components of test taking. *Personnel Psychology, 43,* 695-716.

*Babad, E.Y., & Bashi, J. (1977). Age and coaching effects on the reasoning performance of disadvantaged and advantaged Israeli children. *Journal of Social Psychology, 103,* 169-176.

*Basso, M.R., Carona, F.D., Lowery, N., & Axelrod, B.N. (2002). Practice effects on the WAIS-III across 3- and 6-month intervals. *The Clinical Neuropsychologist, 16,* 57-63.

Best, J.B. (1999). *Cognitive psychology* (5th ed.). Belmont, CA: Brooks/Cole.

Bobko, P. (2001). *Correlation and regression: Applications for industrial/organizational psychology and management* (2nd ed.). Thousand Oaks, CA: Sage.

*Bunting, B.P., & Mooney, E. (2001). The effects of practice and coaching on test results for educational selection at eleven years of age. *Educational Psychology, 21,* 243-253

*Brounstein, P.J., & Holahan, W. (1987). Patterns of change in Scholastic Aptitude Test performance among academically talented adolescents. *Roeper Review, 10,* 110-116.

Cahan, S., & Cohen, N. (1989). Age versus schooling effects on intelligence development. *Child Development, 60,* 1239-1249.

Campbell D.T., & Kenny D.A. (1999). *A primer on regression artifacts.* New York: Guilford Press.

Carretta, T.R., Zelenski, W.E., & Ree, M.J. (2000). Basic Attributes Test (BAT) retest performance. *Military Psychology, 12,* 221-232.

Carroll, J.B. (1993). *Human cognitive abilities: A survey of factor analytic studies.* New York: Cambridge University Press.

*Catron, D.W., & Thompson, C.C. (1979). Test-retest gains in WAIS scores after four retest intervals. *Journal of Clinical Psychology, 35,* 352-357.

Chernyshenko, O.S., & Ones, D.S. (1999). How selective are psychology graduate programs? The effects of the selection ratio on GRE score validity. *Educational and Psychological Measurement, 59,* 951-961.

*Cliffordson, C. (2004). Effects of practice and intellectual growth on performance on the Swedish Scholastic Aptitude Test (SweSAT). *European Journal of Psychological Assessment, 20,* 192-204.

Cohen, J., Cohen, P., West, S.G., & Aiken, L.S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

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

Cook, T.D., & Campbell, D.T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Boston: Houghton Mifflin.

Cortina, J. (2003). Apples and Oranges (and pears, oh my!): The search for moderators in meta-analysis. *Organizational Research Methods, 6,* 415-439.

*Dodrill, C.B. (1983). Long-term reliability of the Wonderlic Personnel Test. *Journal of Consulting and Clinical Psychology, 51,* 316-317.

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

Dunlap, W.P., Cortina, J.M., Vaslow, J.B., & Burke, M.J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods, 1,* 170-177.

*Elwood, D.L., & Clark, C.L. (1978). Computer administration of the Peabody Picture Vocabulary Test to young children. *Behavior Research Methods & Instrumentation, 10,* 43-46.

*Evans, F.R., & Pike, L.W. (1973). The effects of instruction for three mathematics item formats. *Journal of Educational Measurement, 10,* 257-272.

*Ferrer, E., Salthouse, T.A., Steward, W.F., & Schwartz, B.S. (2004). Modeling age and retest processes in longitudinal studies of cognitive abilities. *Psychology and Aging, 19,* 243-259.

*Frankel, E. (1960). Effects of growth, practice, and coaching on Scholastic Aptitude Test scores. *Personnel and Guidance Journal, 38,* 713-719.

*Gavurin, E.I. (1973). Practice effect in anagram solving. *Journal of Psychology, 84,* 279-282.

Gilliland, S.W. (1993). The perceived fairness of selection systems: An organizational justice perspective. *Academy of Management Review, 18*, 694-734.

*Gilmore, M.E. (1927). Coaching for intelligence tests. *Journal of Educational Psychology, 18,* 119-121.

*Goodman, J.T., Streiner, D.L., & Woodward, C.A. (1974). Test-retest reliability of the Shipley Institute of Living scale: Practice effects or random variation. *Psychological Reports, 35,* 351-354.

Gordon, M.E., Slade, L.A., & Schmitt, N. (1986). The "science of the sophomore" revisited: From conjecture to empiricism. *Academy of Management Review, 11,* 191-207.

*Hager, W., & Hasselhorn, M. (1997). Wirkungen der testwiederholung und entwicklungsbedingte leistungssteigerungen bei der durchfűhrung des CFT 1 mit erstklässlern [Retest effects and effects of developmental changes on the CFT 1 score in testing first grade children]. *Zeitschrift fűr Psychologie, 205,* 205-229.

Harcourt Assessment, Inc. (2005). *Miller Analogies Test: Scoring and score reporting*. Retrieved June 5, 2005, from http://www.milleranalogies.com

*Hausknecht, J.P., Halpert, J.A., Harder, B.K., Kuljanin, G., & Moriarty, M. (2005, April). *Issues in repeated testing: Test attitudes and applicant reactions.* Paper presented at the meeting of the Society for Industrial and Organizational Psychology, Los Angeles, CA.

*Hausknecht, J.P., & Howard, M.J. (2004). *Effects of candidate retesting in an employment context.* Paper presented at the meeting of the Academy of Management, New Orleans, LA.

*Hausknecht, J.P., Trevor, C.O., & Farr, J.L. (2002). Retaking ability tests in a selection setting: Implications for practice effects, training performance, and turnover. *Journal of Applied Psychology, 87,* 243-254.

*Holloway, H.D. (1954). Effects of training on the SRA Primary Mental Abilities (primary) and the WISC. *Child Development, 25,* 253-263.

Hunter, J.E., & Schmidt, F.L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA: Sage.

*Koenig, J.A., & Leger, K.F. (1997). Test-taking behaviors and their impact on performance. *Academic Medicine, 72*, S100-S103.

*Kreit, L.H. (1968). The effects of test-taking practice on pupil test performance. *American Educational Research Journal, 5,* 616-625.

Kulik, J.A., Bangert-Drowns, R.L., & Kulik, C.C. (1984). Effectiveness of coaching for aptitude tests. *Psychological Bulletin, 95,* 179-188.

Kulik, J.A., Kulik, C.C., & Bangert, R.L. (1984). Effects of practice on aptitude and achievement test scores. *American Educational Research Journal, 21,* 435-447.

*Lane, R.G., Penn, N.E., & Fischer, R.F. (1966). Miller Analogies Test: A note on permissive retesting. *Journal of Applied Psychology, 50,* 409-411.

*Lievens, F., Buyse, T., & Sackett, P.R. (2005). Retest effects in operational selection settings: Development and test of a framework. *Personnel Psychology*, *58,* 981-1007.

Lipsey, M.W. (2003). Those confounded moderators in meta-analysis: The good, bad, and ugly. *The Annals of the American Academy of Political and Social Science, 587,* 69-81.

*McIntyre, W. (1954). Difficulty of understanding instructions as a factor in coaching and practice effects in intelligence testing. *British Journal of Educational Psychology, 24,* 122-123.

*McRae, H. (1942). The inconstancy of group test I.Q.'s. *British Journal of Educational Psychology, 12,* 59-70.

*Merriman, C. (1927). Coaching for mental tests. *Educational Administration and Supervision, 13,* 59-64.

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

Nathan, J.S., & Camara, W.J. (1998, September). *Score change when retaking the SAT I: Reasoning Test* (Research Note No. RN-05). New York, NY: The College Board.

*Nkaya, I.N., Huteau, M., & Bonnet, J-P. (1994). Retest effect on cognitive performance on the Raven-38 matrices in France and in the Congo. *Perceptual and Motor Skills, 78,* 503-510.

*Oakland, T. (1972). The effects of test-wiseness materials on standardized test performance of preschool disadvantaged children. *Journal of School Psychology, 10,* 355-360.

*Powell, B., & Steelman, L.C. (1983). Equity and the LSAT. *Harvard Educational Review, 53*, 32-44.

Powers, D.E. (1985). The effects of test preparation on the validity of a graduate admission test. *Applied Psychological Measurement, 9,* 179-190.

Powers, D.E. (1986). Relations of test item characteristics to test preparation/test practice effects: A quantitative summary. *Psychological Bulletin, 100,* 67-77.

*Powers, D.E., & Rock, D.A. (1999). Effects of coaching on SAT I: Reasoning test scores. *Journal of Educational Measurement, 36,* 93-118.

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

*Reeve, C.L., & Lam, H. (2004, April). *The relation between practice effects, scale properties, and test-taker characteristics.* Paper presented at the meeting of the Society for Industrial and Organizational Psychology, Chicago, IL.

*Richardson, F., & Robinson, E.S. (1921). Effects of practice upon the scores and predictive value of the Alpha Intelligence Examination. *Journal of Experimental Psychology, 4,* 300-317.

Roth, P.I., Bevier, C.A., Bobko, P., Switzer III, F.S., & Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. *Personnel Psychology, 54,* 297-330.

*Ruiz, R.A., & Krauss, H.H. (1967). Test-retest reliability and practice effect with the Shipley-Institute of Living scale. *Psychological Reports, 20,* 1085-1086.

*Ryan, J.J., Morris, J., Yaffa, S., & Peterson, L. (1981). Test-retest reliability of the Wechsler Memory Scale, Form I. *Journal of Clinical Psychology, 37,* 847-848.

Sackett, P.R., Burris, L.R., & Ryan, A.M. (1989). Coaching and practice effects in personnel selection. In C. Cooper & I. Robertson (Eds.), *International review of industrial and organizational psychology* (pp. 145-183). New York: Wiley.

Sanchez, R.J., Truxillo, D.M., & Bauer, T.N. (2000). Development and examination of an expectancy-based measure of test-taking motivation. *Journal of Applied Psychology, 85,* 739-750.

*Schubert, J. (1967). Effect of training on the performance of the W.I.S.C. 'block design' subtest. *British Journal of Social and Clinical Psychology, 6,* 144-149.

Scullen, S.E., Bergey, P.K., & Aiman-Smith, L. (2005). Forced distribution rating systems and the improvement of workforce potential: A baseline simulation. *Personnel Psychology, 58,* 1-32.

*Shany, M.T., & Biemiller, A. (1995). Assisted reading practice: Effects on performance for poor readers in grades 3 and 4. *Reading Research Quarterly, 30,* 382-395.

Sin, H.P., Farr, J.L., Murphy, K.R., & Hausknecht, J.P. (2004, August). *An investigation of Black-White differences in self-selection and performance in repeated testing.* Paper presented at the meeting of the Academy of Management, New Orleans, LA.

Smither, J.W., London, M., & Reilly, R.R. (2005). Does performance improve following multisource feedback? A theoretical model, meta-analysis, and review of empirical findings. *Personnel Psychology, 58,* 33-66.

Society for Industrial and Organizational Psychology. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). College Park, MD: Author.

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

Steel, P.D., & Kammeyer-Mueller, J.D. (2002). Comparing meta-analytic moderator estimation techniques under realistic conditions. *Journal of Applied Psychology, 87,* 96-111.

*Stricker, L.J. (1984). Test disclosure and retest performance on the SAT. *Applied Psychological Measurement, 8,* 81-87.

*Te Nijenhuis, J., Voskuijl, O.F., & Schijve, N.B. (2001). Practice and coaching on IQ tests: Quite a lot of *g*. *International Journal of Selection and Assessment, 9,* 302-308

*Thomas, P.H., Busciglio, H.H., & Goldenberg, R.J. (2004, April). *An investigation of the effects of applicant retesting on assessment effectiveness: A look at practical implications within U.S. Customs and Border Protection.* Paper presented at the meeting of the Society for Industrial and Organizational Psychology, Chicago, IL.

*Tuma, J.M., Appelbaum, A.S., & Bee, D.E. (1978). Comparability of the WISC and the WISC-R in normal children of divergent socioeconomic backgrounds. *Psychology in the Schools, 15,* 339-346.

*Tuzinski, K.A., Laczo, R.M., & Sackett, P.R. (2005, April). *Impact of response distortion on retaking of cognitive and personality tests.* Paper presented at the meeting of the Society for Industrial and Organizational Psychology, Los Angeles, CA.

U.S. Department of Labor. (1999). *Testing and assessment: An employer's guide to good practices.* Washington, DC: USDOL, Employment and Training Administration.

*van der Reis, A.P. (1963). Is retesting justified in personnel selection? *Psychologia Africana, 10,* 19-30.

Vernon, P.E. (1961). *The structure of human abilities.* London: Methuen.

Vroom, V.H. (1964). *Work and motivation*. Oxford, England: Wiley.

*Whitely, S.E., & Dawis, R.E. (1974). Effects of cognitive intervention on latent ability measured from analogy items. *Journal of Educational Psychology, 66,* 710-717.

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

*Woehlke, A.B., & Wilder, D.H. (1963). Differences in difficulty of forms A and B of the Otis Self-Administering Test of Mental Ability. *Personnel Psychology, 16,* 395-398.

Table 1

Practice Effects for Tests of Cognitive Ability

Test Sequence	*k*	*N*	*d*	95% CI	δ	SD_δ	95% CrI	% Var
Test 1 to Test 2	107	134,436	.24	.20 to .27	.26	.17	-.07 to .60	12
Test 2 to Test 3	16	2,985	.18	.10 to .26	.20	.00	.19 to .21	100
Test 1 to Test 3	15	2,861	.51	.35 to .66	.56	.26	.05 to 1.07	28

Note. *k* = number of independent samples. *N* = total sample size. *d* = mean effect size corrected for sampling error. CI = confidence interval. δ = mean effect size corrected for sampling error and measurement error. CrI = credibility interval. % Var = percentage of variance attributable to artifacts.

Table 2

Means, Standard Deviations, and Correlations of Moderator Variables

Variable	*M*	*SD*	*d*	Study context	Test form	Coaching time
Study context[a]	1.13	.34	-.08			
Test form[b]	1.38	.49	.48**	.10		
Coaching time[c]	3.00	11.43	.32**	-.02	.20*	
Instruction time[d]	44.74	95.81	.12	.37**	-.03	.14

Note. $k = 93$. Values for *d* are observed/uncorrected. [a]Operational contexts coded higher than research settings. [b]Identical forms coded higher than alternate forms. [c]Coaching time represented in hours. [d]Instruction time represented in days.

* $p < .05$, ** $p < .01$.

Table 3

Results of Moderator Analysis of Practice Effects using WLS Regression

Variable	Unstandardized Weights	Standard Error	Standardized Weights
Intercept	.46**	.14	
Study context[a]	-.12	.09	-.20
Test form[b]	.27**	.09	.37
Coaching time[c]	.02*	.01	.26
Instruction time[d]	.01	.01	.17
Multiple *R*	.45		
R^2	.21		
Adjusted R^2	.17		

Note. $k = 93$. Values for *d* are weighted by the inverse of the sampling error variance.

[a]Operational contexts coded higher than research settings. [b]Identical forms coded higher than alternate forms. [c]Coaching time represented in hours. [d]Instruction time represented in days.

* $p < .05$, ** $p < .01$.

Table 4

Results of Moderator Analysis of Practice Effects using Subgroup Analysis

Variable	*k*	*N*	*d*	95% CI	δ	SD_δ	95% CrI	% Var
Study Context								
Research	88	72,641	.22	.17 to .26	.24	.22	-.19 to .66	11
Operational	19	61,795	.27	.22 to .31	.29	.08	.13 to .46	18
Test Form								
Identical	38	7,331	.40	.33 to .48	.46	.17	.12 to .79	49
Alternate	64	91,601	.22	.17 to .27	.24	.19	-.13 to .62	8
Cognitive Ability Dimension								
Verbal	48	53,453	.17	.12 to .22	.19	.15	-.11 to .49	16
Quantitative	24	23,131	.27	.20 to .34	.30	.16	-.02 to .61	16
Analytical	22	28,570	.29	.20 to .38	.32	.21	-.10 to .74	8
Test Coaching								
No	75	81,374	.21	.17 to .26	.24	.18	-.12 to .60	12
Yes	23	2,323	.64	.44 to .83	.70	.41	-.11 to 1.51	23

Note. k = number of independent samples. N = total sample size. d = mean effect size corrected for sampling error. CI = confidence interval. δ = mean effect size corrected for sampling error and measurement error. CrI = credibility interval. % Var = percentage of variance attributable to artifacts.