Journal of Applied Psychology
2011, Vol. 96, No. 5, 941–955

0021-9010/11/$12.00 DOI: 10.1037/a0023562

Can I Retake It? Exploring Subgroup Differences and Criterion-Related Validity in Promotion Retesting

Chad H. Van Iddekinge
Florida State University

Frederick P. Morgeson
Michigan State University

Deidra J. Schleicher and Michael A. Campion
Purdue University

Despite recent interest in the practice of allowing job applicants to retest, surprisingly little is known about how retesting affects 2 of the most critical factors on which staffing procedures are evaluated: subgroup differences and criterion-related validity. We examined these important issues in a sample of internal candidates who completed a job-knowledge test for a within-job promotion. This was a useful context for these questions because we had job-performance data on all candidates ($N = 403$), regardless of whether they passed or failed the promotion test (i.e., there was no direct range restriction). We found that retest effects varied by subgroup, such that females and younger candidates improved more upon retesting than did males and older candidates. There also was some evidence that Black candidates did not improve as much as did candidates from other racial groups. In addition, among candidates who retested, their retest scores were somewhat better predictors of subsequent job performance than were their initial test scores (rs = .38 vs. .27). The overall results suggest that retesting does not negatively affect criterion-related validity and may even enhance it. Furthermore, retesting may reduce the likelihood of adverse impact against some subgroups (e.g., female candidates) but increase the likelihood of adverse impact against other subgroups (e.g., older candidates).

Keywords: personnel selection, practice effects, retesting, demographic differences, criterion-related validity

The practice of allowing applicants to retake staffing procedures has received increased attention among human resources practitioners in recent years (e.g., Bourdeau, 2008; Wheeler, 2004). This is understandable given that decisions about retesting may have substantial consequences for organizations. On the one hand, allowing candidates to retest may increase applicant pools, enhance corporate reputation, and even lead to better selection decisions if retest scores are better indicators of candidates' job-relevant attributes than are their initial test scores. Retesting also may reduce adverse impact and increase workforce diversity if candidates from underrepresented groups tend to improve when given another opportunity.

This article was published Online First April 25, 2011.

Chad H. Van Iddekinge, College of Business, Florida State University; Frederick P. Morgeson, The Eli Broad Graduate School of Management, Michigan State University; Deidra J. Schleicher and Michael A. Campion, Krannert School of Management, Purdue University.

An earlier version of this article was presented at the 70th Annual Meeting of the Academy of Management, Montreal, Quebec, Canada, August 2010.

We thank Philip Roth for his helpful comments on an earlier version of this article.

Correspondence concerning this article should be addressed to Chad H. Van Iddekinge, College of Business, The Florida State University, 821 Academic Way, P.O. Box 3061110, Tallahassee, FL 32306-1110. E-mail: cvanidde@fsu.edu

On the other hand, retesting requires additional time and resources, which may be costly for organizations. Furthermore, if retest scores are not as valid as initial test scores, allowing retesting may result in the selection of less qualified workers and, in turn, reduce organizational effectiveness. Of course, whether organizations offer retesting also can have important consequences for the livelihood and well-being of job candidates, who may want or need a particular job but who initially are not selected.

This increased applied interest in retesting has stimulated research that attempts to better understand the nature and implications of retest effects. Most studies to date have examined how students' performance on cognitive-oriented tests—completed for research purposes or for college admission—changes with retesting (e.g., Lievens, Buyse, & Sackett, 2005; Lievens, Reeve, & Heggestad, 2007; Reeve & Lam, 2007; te Nijenhuis, Voskuijl, & Schijve, 2001). This research has shown that, in general, students tend to obtain higher scores upon retesting.

In contrast, the few studies that have examined the effects of repeat testing in employment contexts have tended to focus on personality measures. Results of this work have been inconsistent, with some studies finding that applicants or current employees tend to obtain higher scores upon retesting (e.g., Hausknecht, 2010; Kelley, Jacobs, & Farr, 1994) and other studies finding small or no score differences upon retesting (e.g., Ellingson, Sackett, & Connelly, 2007; Hogan, Barrett, & Hogan, 2007).

Although the results of this research have contributed greatly to the understanding of retest effects, important questions remain.

The purpose of the current study was to begin to address some of these questions and to extend prior research in four key ways. First, most retesting research has focused on how mean test scores change with retesting (Hausknecht, Halpert, Di Paolo, & Moriarty Gerrard, 2007). Although this is important, the effects of retesting on other outcomes are equally, if not more, important to practitioners. One critical factor on which staffing procedures are evaluated is subgroup differences, which are a precursor of adverse impact. We examined whether there are subgroup differences in retest effects. The existence of such differences is important because if certain candidates (e.g., females) tend to gain more from retesting than other candidates (e.g., males), inclusion of retesters could influence subgroup differences and, in turn, reduce or exacerbate adverse impact. A better understanding of subgroup differences in retest effects also has implications for theory and research. For instance, finding that certain types of people tend to improve more (or not) with retesting than other people may enrich theories of individual differences.

Another critical factor on which staffing procedures are evaluated is criterion-related validity. The validity of inferences drawn from scores on staffing procedures has implications for selecting and promoting individuals who are most likely to have high job performance and contribute to organizational success. Unfortunately, little is known about how retesting affects criterion-related validity in general and how it affects validity in employment contexts in particular.

Thus, a second way we have extended previous research is by providing some of the first data on how retesting affects the prediction of job performance. This focus is important for several reasons. For example, if retesting changes the constructs a given test measures (e.g., Lievens et al., 2007), this could increase or decrease the criterion-related validity of the test and, in turn, the quality of candidates selected on the basis of the test. The effects of retesting on validity also could have implications for the conduct and interpretation of studies that attempt to estimate or compare the validity of various staffing procedures. For instance, sample composition (i.e., one-time applicants, repeat applicants, or both types of applicants) could represent an important source of variance in validity estimates across primary studies within meta-analytic investigations of staffing procedures.

Third, the limited field research on retesting has been conducted primarily with job applicants who completed tests during the selection process (e.g., Hausknecht, Trevor, & Farr, 2002; Hogan et al., 2007). In applicant samples, criterion data are available only for those who ultimately are selected and who remain on the job long enough for their performance to be evaluated. This situation represents one of the main challenges for determining whether and how retesting affects the criterion-related validity of the selection procedures. Specifically, using the test(s) as a basis for selection produces a restricted range of predictor scores, which can lead to biased estimates of validity (Hunter, Schmidt, & Le, 2006). Not having criterion data on applicants who failed the selection test also reduces the available sample from which validity can be estimated.

Participants in the current study completed a test for a within-job promotion. That is, once promoted, employees continued to perform the same job; the only difference was that promoted employees had the authority to make certain decisions on their own, whereas employees who were not promoted had to get supervisor approval. This provided a useful context for understanding the effects of retesting on criterion validity because candidates remained in the job, regardless of whether they passed or failed the promotion test. Thus, we were able to obtain postdecision criterion data on all candidates. This enabled us to estimate validity in the absence of (direct) range restriction and to conduct the analyses on a relatively large sample of initial and retest candidates. Because range restriction also can bias estimates of mean differences (Roth, Van Iddekinge, Huffcutt, Eidson, & Bobko, 2002), this context also allowed us to more clearly understand the nature and effects of subgroup differences in retesting.

Finally, because most studies have examined retest effects for cognitive ability or personality tests, relatively little is known about retest effects for other staffing procedures. Accordingly, we examined retesting on a job-knowledge test. This is important because knowledge-based tests are widely used to facilitate staffing decisions and have been shown to be among the best predictors of job performance (Schmidt & Hunter, 1998). In addition, whereas cognitive ability and personality are considered to be relatively stable constructs, job knowledge tends to be more malleable and can be increased through learning and experience. Thus, the factors that influence retest effects on knowledge-based tests (e.g., learning) may be quite different from the factors that influence retest effects on cognitive ability tests (e.g., test-taking strategies) and personality tests (e.g., response distortion). In this regard, we also explore whether and how the time span between initial test and retest occasions influences retesting effects.

Effects of Retesting on Test Performance

Retest effects have been defined as score changes that occur after prior exposure to an identical or alternate form of a test under standardized conditions (Lievens et al., 2005). As noted, most research has examined whether people improve their scores on cognitive-oriented tests upon retesting. Hausknecht et al. (2007) cumulated evidence from this research base ($k = 107$) and found an overall standardized mean difference (d) of .26. This suggests that retest scores tend to be about one quarter of a standard deviation higher than initial test scores.

Far fewer studies have examined retest effects on knowledge-oriented measures. Lievens et al. (2005) reported a retest effect of .27 for scores on a test of science knowledge used to assess medical school applicants. Raymond, Neustel, and Anderson (2007) found retest effects of .79 and .48 on two certification tests completed by medical imaging workers. Most recently, Schleicher, Van Iddekinge, Morgeson, and Campion (2010) reported a retest effect of .15 on a job-knowledge test used to select applicants for professional jobs within a federal agency.

Lievens et al. (2007) delineated a framework of factors that may lead to score increases with retesting. Two factors from this framework provide a basis to predict that candidates who complete a job-knowledge test will score higher upon retesting. First, unlike staffing procedures that measure relatively stable constructs such as cognitive ability and personality, job knowledge is malleable, and thus, test-score improvement may reflect true changes in knowledge. For example, initial exposure to the test may lead to true increases in job knowledge upon retesting. This explanation is consistent with the *testing effect*, whereby the act of taking a test

not only assesses what candidates know but also increases their knowledge of the subject (Roediger & Karpicke, 2006).

True changes in job knowledge also may occur as a result of the passage of time due to general learning and experience or to active steps candidates may take, such as studying test content or seeking relevant training and experience. This possibility is particularly relevant in promotion contexts, because candidates who initially fail remain in the job (i.e., in a within-job promotion context) or remain in a different job within the organization (i.e., in a between-job promotion context). Thus, these individuals have an opportunity to gain additional knowledge and experience by performing the job, or at least by continuing to work within the organization.

A second reason why candidates may improve on job-knowledge tests is that retesting reduces "debilitating, construct-irrelevant influences that are present (or at least more salient) during the initial testing session" (Lievens et al., 2007, p. 1675). For example, some candidates may experience confusion or anxiety when taking a test for the first time, and there is evidence that anxiety tends to negatively affect test performance (e.g., Cassady & Johnson, 2002; Hausknecht, Day, & Thomas, 2004; McCarthy & Goffin, 2005). Given prior exposure to the test, candidates may experience less confusion and anxiety upon retesting because they are more familiar with the testing environment, test format, and so forth. This increased familiarity and reduced anxiety may, in turn, enable candidates to perform better on the retest.

Finally, in addition to these more substantive reasons for score increases, retest scores may be higher as a result of regression to the mean. Regression to the mean can occur when the same construct is measured on two or more occasions and the measure from the first occasion yields extreme scores on the construct. The result is that extreme scores from the first occasion will tend to regress (upward or downward) toward the mean on the second occasion (Bobko, 2001). In retesting situations, candidates who retest may represent an extreme group (of low scores) because their initial test scores were not high enough to be selected. Thus, scores of some portion of retesters should regress (upward) to the mean upon retesting. For example, regression effects may occur when illness or other transient factors negatively affect how candidates perform on the initial test (Hausknecht et al., 2007). For this and the other reasons described above, we predicted the following:

> *Hypothesis 1:* Candidates will improve their job-knowledge test performance upon retesting.

Although research by Lievens et al. (2007) and others has delineated factors that may contribute to retesting effects, very few studies have attempted to measure these factors. One likely reason for this is that it can be quite difficult to collect accurate data from job candidates regarding their true levels of anxiety, recall of test content, test preparation, and so forth. One way to begin to assess such factors is by examining relations between retest-score change and the amount of time between initial test and retest occasions. The various underlying causes of retest effects may produce different patterns of relationships between time lag and potential score improvement. For example, if score improvement is due to candidates' recall of specific items on the test and how they initially responded to them, there may be negative relationship between time lag and score change. Conversely, if score improvement is due to reduced anxiety, we may not necessarily expect any relationship between time lag and score change. Finally, if score improvement is due to real changes in the target construct, such as from learning and experience or from candidates' preparation for the retest, there may be a positive relationship between time lag and score change. Given the reasons we noted for why job-knowledge test scores are likely to increase because of real gains in job knowledge (e.g., due to additional experience on the job), we expected to find support for this last possibility.

> *Hypothesis 2:* There will be a positive relationship between time lag and changes in performance on the job-knowledge test, such that the longer the time between initial and retest occasions, the more candidates will improve upon retesting.

Subgroup Differences in Retesting Effects

We know of only one previous study that investigated subgroup differences in retest effects. Schleicher et al. (2010) explored this issue across eight selection procedures, including written measures of cognitive ability, job knowledge, and biodata, as well as a structured interview and two assessment center exercises. Results revealed that White applicants showed larger score improvements than Black and Hispanic applicants on all three written measures. In contrast, subgroup differences in retesting were smaller and in some cases favored minority applicants on the structured interview and assessment center. Furthermore, females tended to improve more than males, particularly on the structured interview and assessment center. Regarding age, applicants under 40 years old showed larger score improvements than did applicants 40 and older on all eight selection procedures. We build upon the initial work of Schleicher and colleagues by examining race, gender, and age differences in retest effects on a job-knowledge test in a promotion context.

Race Differences

Theory and empirical research suggest that White candidates may improve more upon retesting than racial minorities, particularly Blacks and Hispanics. First, it is well documented that Whites tend to score higher on measures of mental ability than do Blacks and Hispanics (Roth, BeVier, Bobko, Switzer, & Tyler, 2001). In turn, there is research to suggest that ability affects how much people improve with retesting (e.g., Kulik, Kulik, & Bangert, 1984; Rapport, Brines, Axelrod, & Theisen, 1997; Vernon, 1954). For example, Kulik et al.'s (1984) meta-analysis revealed mean retest effects (d) of .17, .40, and .82 (respectively) on various standardized aptitude and achievement tests among students who possessed low, medium, and high levels of ability.

High-ability job candidates may learn more from the initial test-taking experience than low-ability candidates. For example, high-ability candidates may remember more about the initial test, such as the test content and how they responded to certain questions, and then use this information to prepare for, and respond to, questions on the retest. Consistent with this possibility, Lievens et al. (2007) found evidence that retest scores correlated significantly with scores on a memory association test ($r = .29$), whereas initial test scores did not ($r = -.03$). The idea that ability may be related to retest performance is particularly relevant to measures of job-

knowledge tests because cognitive ability is a direct precursor of knowledge acquisition (Schmidt, Hunter, & Outerbridge, 1986).[1] Thus, not only may high-ability candidates learn more from the initial test than low-ability candidates, they also may be better able to acquire additional job knowledge, and do so faster, to improve their scores upon retesting than low-ability candidates.

A second reason why White candidates may improve more upon retesting concerns differences in test-taking attitudes and motivation, which can influence performance on employment tests (e.g., Arvey, Strickland, Drauden, & Martin, 1990; Bauer, Maertz, Dolen, & Campion, 1998; Gilliland, 1993). Specifically, there is some evidence that White applicants tend to hold more positive testing attitudes and motivation than do minority applicants (e.g., Arvey et al., 1990; Chan & Schmitt, 1997; Chan, Schmitt, Sacco, & DeShon, 1998). If attitudes and motivation about testing influence initial test scores, it is possible that people who tend to hold more favorable attitudes (e.g., Whites) also may gain more from practice than people with less favorable attitudes (e.g., minorities). For example, it has been suggested that whereas Whites tend to believe test scores reflect a person's competence, Blacks are more likely to believe test performance is due to factors outside their control, such as luck or test-taking conditions (Chan, 1997; Helms, 1992). As such, White candidates, who are more likely to believe they can improve their test scores, may devote more time and effort to preparing for a retest than minority candidates.

At least two studies have found some support for the influence of test-taking attitudes on score changes with retesting. Reeve and Lam (2007) found some evidence that constructs such as test-taking self-efficacy and a general belief in tests related positively to score gains on a cognitive ability test within a sample of undergraduate test-takers. Similarly, Chan, Schmitt, DeShon, Clause, and Delbridge (1997) reported that test-taking motivation affected undergraduates' performance on a parallel set of cognitive tests, controlling for the effects of performance on the first set of tests.

Thus, we predicted that White candidates would improve more upon retesting than would Black and Hispanic candidates. We also explored possible differences between White and Asian candidates, but we did not have specific expectations regarding the nature or magnitude of retest differences between these two groups.

> *Hypothesis 3:* White candidates will improve more on the job-knowledge test upon retesting than Black and Hispanic candidates.

Gender Differences

Although staffing researchers tend to focus on racial group differences, gender differences also are important and are becoming increasingly so given the changing nature of today's workforce. For instance, women now comprise approximately one half of the U.S. workforce (U.S. Bureau of Labor Statistics, 2009).

There are several reasons to expect that female candidates will improve more upon retesting than male candidates. First, research has shown that compared to males, females tend to react more positively to negative feedback (e.g., failing a staffing test) and also are more likely to make use of that feedback (Johnson & Helgeson, 2002). For example, research by Roberts and Nolen-Hoeksema (1989, 1994) suggested that, compared to males, females perceive feedback to be more accurate and to provide more useful information about themselves. Other research has found that female employees are more likely to comply with supervisors' recommendations concerning performance improvement than are male employees (Sachau, Houlihan, & Gilbertson, 1999).

Second, there is research to suggest that females are somewhat more likely than males to attribute failure to internal factors than to external factors (e.g., Boggiano & Barrett, 1991; Hirschy & Morris, 2002). This suggests that the two genders may assign different attributions to their initial failure on an employment test, which, in turn, may influence how they approach subsequent test attempts. Specifically, because females tend to attribute low test performance to something within themselves (e.g., inadequate preparation, lack of knowledge), they may be more likely to attempt to rectify the situation, such as by preparing more for a retest. In contrast, because males tend to attribute failure to factors outside their control (e.g., bad luck, a poor test), they may be less likely to try to make changes prior to a retest. This reasoning is consistent with research by Fisher, Harrison, and Nadler (1978), who found that after receiving help on a difficult task, males tended not to consult a self-help booklet, even though they were told that doing so would lead to better performance on a subsequent task.

Finally, research has shown that females tend to be somewhat more anxious than males (e.g., Duehr, Jackson, & Ones, 2004; Feingold, 1994; Macoby & Jacklin, 1974). As noted, anxiety tends to reduce performance on employment tests, and allowing applicants to retest is thought to reduce test-taking anxiety (e.g., due to greater familiarity with the testing situation; Lievens et al., 2007). Therefore, females may have more to gain (i.e., in terms of reduced anxiety) than males given the opportunity to retest. On the basis of these arguments, we predicted the following:

> *Hypothesis 4:* Female candidates will improve more on the job-knowledge test upon retesting than male candidates.

Age Differences

We also investigated the relationship between age and retest effects, which is timely given current employment trends. For example, it is estimated that by 2016, 44% the U.S. workforce will be 45 years old or older (U.S. Bureau of Labor Statistics, 2009). The long-term trend toward earlier retirement also has been reversed in the past few years. In fact, as many as three out of four older workers plan to launch a second career after retirement (Greene, 2005).

Younger candidates are likely to improve more with retesting than older candidates due to age differences in goal orientation and

[1] The idea that preexisting differences in mental ability affect subgroup differences in retest effects may not be relevant to all types of selection procedures. For example, research that has examined retest effects on mental ability tests has found that score gains are inversely related to the *g* loading of the tests (e.g., te Nijenhuis, van Vianen, & van der Flier, 2007). This suggests that retest effects on such tests are due to factors other than initial levels of general intelligence, including test-wiseness or narrow abilities such as memory association (Jensen, 1998; Lievens et al., 2007; te Nijenhuis et al., 2007).

motivation. For example, selection, optimization, and compensation (SOC) theory (Baltes & Baltes, 1990; Freund & Baltes, 2000) describes how motivational processes focus on either optimizing gains or minimizing losses and how peoples' motivations tend to change over the life span. Optimizing processes include identification and commitment to goals aimed at higher levels of performance, as well as the acquisition of new knowledge and skills for achieving those goals. In contrast, loss-minimizing processes involve adjusting one's goals or redirecting resources in an effort to avoid actual or impending losses. Thus, although both processes involve the investment of time and resources to achieve particular goals, they differ in regard to their focus on achieving higher levels of functioning (i.e., optimization) versus counteracting losses (i.e., compensation).

Research on SOC theory and goal orientation consistently has shown that younger people tend to focus on growth and performance optimization, whereas older people tend to focus on avoiding losses (e.g., Ebner, Freund, & Baltes, 2006; Freund, 2006; Heckhausen, 1997; Hyvönen, Feldt, Salmela-Aro, Kinnunen, & Makikangas, 2009). This finding is in line with Kanfer and Ackerman's (2004) propositions regarding age and motivation. They suggested that for younger workers, there tends to be a high level of utility for increasing work performance, which can lead to valued outcomes such as pay increases, recognition, and promotion. However, as workers age, achievement motives decrease, and motives related to positive affect and protecting one's self-concept increase. For instance, compared to younger workers, older workers may be less motivated to obtain another promotion because occupational achievement plays a relatively smaller role in their lives (Kanfer & Ackerman, 2004).

Thus, when younger candidates fail an initial test, they may be quite motivated to pass the retest, because doing so will help them achieve optimization-oriented goals, such as obtaining a promotion, as was the case in the current study. As such, younger candidates may be willing to invest the necessary resources (e.g., time and effort to prepare) to pass the test when retaking it. On the other hand, although older candidates may choose to retest, they may be less willing to invest additional resources to enhance their test performance, particularly if they believe doing so will require a substantial level of effort and create losses in other areas of work or life in general (Kanfer & Ackerman, 2004).

A second reason why younger candidates may improve more upon retesting than older candidates is due to age differences in fluid intelligence (Cattell, 1987), which is associated with working memory, abstract reasoning, and processing new information (Kanfer & Ackerman, 2004). In contrast to crystalized intelligence, which does not tend to decline with age, fluid intelligence peaks in adolescence and begins to decline progressively in one's 30s (Horn & Cattell, 1967; Kanfer & Ackerman, 2004; Schaie, 1996). This is important because fluid intelligence influences how quickly people can acquire new knowledge. Thus, when the time between test occasions is relatively short (as was the case in the present study), younger candidates may have the ability to gain more knowledge than older candidates and, in turn, improve more upon retesting. On the basis of these arguments, we predicted the following:

> *Hypothesis 5:* Younger candidates will improve more on the job-knowledge test upon retesting than older candidates.

Effects of Retesting on Criterion-Related Validity

Several studies have explored the effects of retesting on criterion-related validity in academic settings (e.g., Allalouf & Ben-Shakhar, 1998; Coyle, 2006; Lievens et al., 2005, 2007; Reeve & Lam, 2005). This research has found that retest scores tend to be better predictors of criteria than initial test scores, although initial–retest validity differences often are small. For example, Allalouf and Ben-Shakhar (1998) found that retest scores on a mental ability test correlated more highly with scores on a college matriculation exam than did initial test scores, particularly for the verbal portion of the test (e.g., *r*s = .56 vs. .46 in the uncoached control group). Coyle (2006) reported that initial and retest scores on the SAT and ACT correlated .50 and .54, respectively, with college grade point average (GPA; although the two validities are not significantly different). Of particular relevance to the present study, Lievens et al. (2005) found that medical school applicants' retest scores on a science-knowledge test were significantly more predictive of subsequent GPA than were their initial test scores (*r*s = .21 vs. .11).

We know of only one study that examined retesting and criterion-related validity in an employment setting. In a sample of law enforcement applicants, Hausknecht et al. (2002) found that initial and retest scores on a battery of selection tests did not differ significantly in their prediction of subsequent training exam scores and turnover. However, there was some evidence of a trend that retest scores on a cognitive ability test were better predictors of the criteria (e.g., *r*s = .31 and .27, respectively, with training exam scores).

We built upon the above research by investigating whether and how retesting affects the prediction of job performance. We again drew on Lievens et al.'s (2007) framework, which suggests three possible ways in which retesting may affect the criterion validity of a job-knowledge test. The first possibility is no criterion-related validity differences between initial and retest scores. If score changes from the initial test to retest reflect only actual changes in the underlying construct (e.g., increases in job knowledge due to initial exposure to the test), then scores on the retest may capture individual differences in job knowledge to an extent similar to that of scores on the initial test. In other words, the relative order of candidates' test scores may remain largely intact, which would result in little or no validity difference between initial and retest scores as related to job performance.

The second possibility is that initial test scores are better predictors of job performance than retest scores. Score changes from the initial test to the retest may be influenced by test-specific knowledge or test-taking skills gained by initial exposure to the test. This phenomenon often is referred to as test-wiseness (Millman, Bishop, & Ebel, 1965). For instance, after taking the initial test, candidates may gain understanding for how to progress through the test, narrow their choices, and structure their time. Importantly, test-taking skills such as these are independent of candidates' knowledge of the test content. To the extent this occurs, initial test scores may be better indicators of job knowledge than retest scores, which are more likely to reflect test-wiseness. As such, candidates' initial scores may demonstrate stronger evidence of criterion-related validity with job performance.

The final possibility is that retest scores are better predictors of job performance than initial test scores. This may occur for two

reasons. First, retest scores may be better indicators of the underlying construct due to a reduction in the influence of construct-irrelevant factors, such as test anxiety. Second, there may be greater variance in retest scores than in initial test scores, which would increase the potential magnitude of relationships between retest scores and other variables, including measures of job performance. For example, earlier we noted possible individual differences in how much candidates learn from the initial test and their ability to gain additional knowledge prior to the retest. We also noted how there may be individual differences in motivation to pass the retest. Some candidates may prepare more for the retest, which may result in greater variance in job knowledge upon retesting. The possibility of greater variance in retest scores than in initial test scores would be consistent with past retesting research (e.g., Schleicher et al., 2010), as well as with research showing that training can increase preexisting individual differences among trainees (e.g., Alliger & Katzman, 1997; McGehee & Thayer, 1961).

We expected to find support for this latter explanation, namely, that retest scores will be better predictors of job performance than initial test scores. Retest scores may be better indicators of job knowledge because construct-irrelevant factors, such as test anxiety and unfamiliarity, should play less of a role upon retesting. Furthermore, although both initial and retest scores should capture individual differences in job knowledge, retest scores may reflect a larger range of job knowledge due to individual differences in how much candidates learn from the initial test and prepare for the retest. In contrast, because most job-knowledge tests are objective (i.e., answers are either correct or incorrect), retest scores on such tests would seem less likely to capture extraneous factors, such as test-taking tricks and response distortion, which would reduce the validity of retest scores.

Hypothesis 6: Retest scores on the job-knowledge test will be more predictive of job performance than initial test scores.

Method

Sample and Design

The data were collected during a validation study for a new within-job promotion process in a government organization. The job of interest involved reviewing source materials (e.g., descriptions, diagrams, pictures, results of research) in the area of science and technology to evaluate and determine the value of a product or method. The job-knowledge test, which measured knowledge of legal issues associated with the introduction of new technology (see below), represented the sole basis for promotion decisions. Once promoted, employees continued to perform the same job (i.e., a within-job promotion). The only difference was that promoted employees had the authority to make final decisions about the products and methods, whereas employees who were not promoted had to first seek supervisor approval concerning these decisions. However, making these decisions was not a daily activity, and the knowledge, skills, abilities, and other characteristics and behaviors related to decision autonomy were not assessed in either the promotion test or in the job-performance criterion.

A total of 605 candidates initially completed the promotion test, of whom 330 passed and 275 failed (pass rate = 54.5%). Of the 275 candidates who failed, 192 chose to retake the test at a later date. Thus, almost 70% of candidates who initially failed chose to retest. The organization offered the test several times a year, and employees could choose to retake the test at any time. The time lag between administrations of the initial test and retest ranged from 0.63 months to 18.7 months, with a mean of 3.93 months. Approximately 3 months after all the testing was completed (including retesting), job-performance information was collected from candidates who took the promotion test, regardless of whether they passed or failed the test.

Table 1 presents demographic information for the study sample. Participants were ethnically diverse, with Black, Asian, and Hispanic candidates comprising 58.4% of the sample (the remaining 41.6% of the sample was White). In terms of gender and age, 71.2% of participants were male, and 72.4% were under the age of 40, with a mean age of 35.5 years (SD = 9.96).

Promotion Test

The promotion test was developed from a test blueprint that specified the number of test questions for each legal knowledge area based on a comprehensive job analysis and linkage ratings of subject matter experts (SMEs). All items were multiple-choice format with one correct answer and three distractors. An initial version of the test was piloted on a sample of 300 supervisors. The data from this study were used to evaluate the quality of the test items (e.g., item difficulty, item–total correlations). The final set of items was used to create eight forms of the test. Each form comprised 50 items (from a bank of several hundred items) that were strictly linked to the test blueprint and were judged to be content valid according to the SMEs (mean α = .83). The process used to determine which form candidates received, for both the initial test and the retest, was essentially random.

Job-Performance Measure

Supervisors were asked to evaluate the performance of each promotion candidate, and we were able to obtain performance

Table 1
Demographic Characteristics of Study Participants

	Initial test		Retest	
Subgroup	*N*	%	*N*	%
Race				
Asian	232	38.4	88	46.1
Black	80	13.3	34	17.8
Hispanic	41	6.8	18	9.4
White	251	41.6	51	26.7
Gender				
Male	431	71.2	139	72.4
Female	174	28.8	53	27.6
Age				
20s	217	37.0	31	16.3
30s	202	34.5	70	36.8
40s	108	18.4	59	31.1
50s	40	6.8	21	11.1
60s	19	3.2	9	4.7
Overall	605		192	

Note. There was missing race information for one candidate and missing age information for 19 candidates.

ratings for 403 of the 605 candidates. Supervisors did not have access to candidates' test scores and were informed that their ratings were for research only and would not be shared with the candidates. As noted, successful and unsuccessful candidates performed the same job. The only difference was that promoted candidates did not have to seek supervisor approval for certain decisions. Thus, although we used a predictive design, the criterion was concurrent in the sense that the job performance of all participants was rated against the same criteria, regardless of whether they were promoted.

The performance of each participant was rated on three dimensions: job knowledge (i.e., the amount of job-relevant information employees possess), skill (i.e., the proficiency with which employees perform their job), and overall performance (i.e., everything considered, how well employees perform their job). Each dimension was rated on a 7-point Likert-type scale, with anchors that ranged from 1 (*well below average*) to 7 (*well above average*). Because supervisors' ratings of the three dimensions were highly correlated, we averaged the dimension ratings to form a unit-weighted composite measure of candidates' job performance ($\alpha = .97$).

For 152 of the candidates, we were able to collect ratings on the same dimensions from a second supervisor who was familiar with a given candidate's performance. This allowed us to estimate the interrater reliability of the performance ratings. We computed intraclass correlation coefficients (ICCs) to estimate reliability based on a single rater (ICC,1) and the reliability of the mean ratings based on two raters (ICC,2). The resulting estimates were .78 and .88, respectively, which suggest a high level of consistency between two supervisors' ratings of the same candidates. Therefore, we used the average of the mean ratings from the two supervisors whenever available.

Results

Analysis of Potential Test-Form Differences and Effects

Before testing our hypotheses, we investigated potential differences among the different promotion test forms and whether such differences affected the results of our substantive analyses. Mean scores on the initial test ranged from 24.8 to 29.27 across forms ($SD = 1.48$). An analysis of variance indicated these mean differences were significant, $F(7, 184) = 2.37$, $p < .05$. Mean scores on the retest ranged from 28.5 to 35.3 across the eight test forms ($SD = 2.14$), and these differences also were significant, $F(7, 184) = 3.13$, $p < .05$. However, post hoc comparisons (with Bonferroni correction) suggested that only one out of the 56 possible pairs of forms was significantly different. Thus, for the most part, mean differences between test forms were small and statically nonsignificant.

To examine whether test form affected relations between promotion test scores and job performance (i.e., criterion-related validity), we set up an analysis of covariance (ANCOVA) model in which test form was a fixed independent variable, test scores were a covariate, and job performance was the dependent variable. We assessed this model twice—once using initial test scores and once using retest scores. The interaction between test form and test scores was nonsignificant both for initial test scores, $F(7, 120) = 0.75$, $p = .61$, and for retest scores, $F(7, 120) = 0.64$, $p = .72$. This suggests that the relationship between test scores and job performance did not depend on the form of the test that candidates completed.

We also investigated the potential influence of the particular sequence of test forms candidates completed (e.g., Form 1 and then Form 8 vs. Form 8 then Form 1). To examine whether test-form sequence affected relations between initial and retest scores, we developed an ANCOVA model in which form sequence was a fixed independent variable, initial test scores were a covariate, and retest scores were the dependent variable. Results revealed that only initial test scores were significantly related to retest scores, $F(1, 136) = 8.28$, $p < .05$. Neither the test-form sequence main effect, $F(26, 136) = 1.01$, $p = .46$, nor the interaction between form sequence and initial scores, $F(26, 136) = 1.00$, $p = .46$, was significant. This suggests that the relationship between initial and retest scores did not depend on the sequence of test forms.

Finally, we examined whether test-form sequence affected relations between candidate demographics and retest effects. We set up the same ANCOVA model as above, except we added gender, age, or race (in separate models) as a fixed independent variable and interacted test-form sequence with the demographic variable. In no instance was the Form Sequence $\times$ Demographic interaction significant. This suggests that the relationship between demographics and retest scores (controlling for initial test scores) did not depend on the sequence of test forms.

Thus, although there were some mean differences across different forms of the job-knowledge test, results of the above analyses suggest that neither the test form nor the sequence in which candidates took the different forms had a significant influence on retest effects, subgroup differences in retest effects, or the criterion-related validity of initial and retest scores. We therefore used data from all the forms to test our hypotheses.

Effects of Retesting on Test Performance

Table 2 presents correlations among the study variables. Table 3 presents descriptive statistics and retest effects for the 192 candidates who initially failed the promotion test and retook it at a later date. Before discussing the results of our hypothesis testing, we first note two other interesting findings. First, the overall correlation between initial and retest scores was .48 (see Table 3), which suggests the existence of individual differences in test-score improvement. In fact, although most candidates improved their test scores upon retesting, a notable proportion (17.7%) obtained lower scores upon retesting. This finding has implications for possible differences between initial and retest scores with respect to criterion-related validity. Specifically, if there were little or no change in the relative order of candidates based on initial and retest scores, then there likely would be little or no difference in criterion-related validity across time. The fact that there were notable relative order changes allows for the possibility that initial and retest scores may differ in the extent to which they predict job performance.

Second, as we expected, the variance in test scores was larger for the retest than for the initial test (overall $SD = 6.68$ vs. 4.98; see Table 3). This occurred despite the fact that test scores overall increased upon retesting, which would tend to decrease score variance (i.e., because most scores are clustered in the upper end

Table 2
Correlations Among the Study Variables

Variable	*N*	1	2	3	4	5	6	7	8	9	10
1. Asian	604	—									
2. Black	604	—	—								
3. Hispanic	604	—	—	—							
4. White	604	—	—	—	—						
5. Gender	605	.01	−.11*	.00	.06	—					
6. Age	586	.18*	−.04	−.05	−.13*	.14*	—				
7. Initial test scores	605	−.15*	−.16*	−.06	.30*	−.03	−.38*	—			
8. Retest scores	192	.02	−.10	−.06	.12	−.08	−.23*	.48*	—		
9. Time lag	192	−.10	.04	.04	.06	−.08	.01	−.22*	.00	—	
10. Job performance	403	−.03	−.22*	−.07	.24*	−.05	−.19*	.48*	.38*	−.02	—

Note. Demographic variables are coded 1 = target group (e.g., Asian) and 0 = all other groups. Gender is coded 1 = male and 0 = female. Age is coded as continuous. Time lag reflects days between initial test date and retest date.
* $p < .05$, two-tailed.

of the distribution). This finding suggests that rather than reducing scores differences among candidates (e.g., because all candidates now know what the test involves), initial exposure to the test (and what candidates may do as a result of failing the test) actually increases score differences among candidates.

Hypothesis 1 predicted that candidates would improve their job-knowledge test performance upon retesting. Candidates overall improved their scores by an average of .93 standard deviations, which equates to an increase of 5.43 points on the test. The average increase in scores from the initial test to the retest was statistically significant, $t(191) = 12.25$, $p < .05$. This provides support for Hypothesis 1.

As discussed, some portion of retest effects may be due to regression toward the mean. We estimated the amount of regression effects using the approach described by Bobko (2001, pp. 162–167). This approach requires the correlation between initial and retest scores ($r = .48$) and the difference between the mean score of retesters on the initial test and the mean of all candidates on the initial test. Because the initial–retest score correlation is subject to range restriction due to selection on the initial test (i.e., only candidates who failed the initial test have retest scores), we first corrected this correlation for direct range restriction using Thorndike's (1949) Case II formula, which yielded a value of .66. Results suggested that .41 of the .93 retest effect (44%) can be attributed to regression to the mean, whereas .52 of this effect (56%) cannot be attributed to regression to the mean. This result is very consistent with Raymond et al. (2007), who found that about one half of score increases on a certification test were due to regression effects.

Hypothesis 2 predicted there would be a positive relationship between time lag and score change on the job-knowledge test, such that the longer the time between initial and retest occasions, the more candidates would improve upon retesting. We used multiple regression analysis to test this hypothesis, whereby retest scores were regressed onto initial test scores and a variable that reflected the number of days between each candidate's initial test and retest. A statistically significant beta weight for the latter variable would

Table 3
Descriptive Statistics and Retest Effects for Promotion Test Scores by Subgroup and Overall

		Initial test		Retest			
Subgroup	*N*	*M*	*SD*	*M*	*SD*	*r*	*d*
Race							
Asian	88	27.09	5.43	33.10	7.28	.48	0.95
Black	34	26.97	4.61	31.50	6.54	.42	0.81
Hispanic	18	26.94	4.47	31.72	4.87	.41	1.02
White	51	28.88	4.44	34.25	5.98	.48	1.03
Gender							
Female	53	26.45	5.48	33.81	6.49	.76	1.23
Male	139	27.91	4.74	32.60	6.74	.38	0.82
Age							
20s	31	27.10	5.71	34.55	6.44	.74	1.23
30s	70	27.81	4.81	34.27	6.63	.40	1.13
40s	59	27.07	4.99	31.56	7.06	.48	0.75
50s	21	27.24	4.30	30.86	5.69	.48	0.72
60s	9	29.00	5.55	31.33	5.74	.43	0.41
Overall	192	27.51	4.98	32.94	6.68	.48	0.93

Note. Race information was missing from one candidate, and age information was missing from two candidates. *r* reflects the zero-order correlation between scores on the two occasions of interest. *d* = standardized mean difference ($[M_{\text{Time 2}} - M_{\text{Time 1}}]/SD_{\text{pooled}}$). All correlations and mean differences are statistically significant ($p < .05$, two-tailed).

provide evidence that retest effects differ by the amount of time between test occasions.

In support of this hypothesis, there was a small, positive relationship ($\beta = .11$, $p < .05$) between time lag and retest performance. That is, the longer the time between candidates' initial test and retest, the more they tended to improve. This finding is consistent with the explanation that test-score improvement was due to real changes in job knowledge (e.g., from additional experience on the job or from time candidates devoted to preparing for the retest), rather than to candidates' recall of specific test content and their initial responses.[2]

Subgroup Differences in Retest Effects

The next set of hypotheses concerned differences in retest effects across candidate subgroups. We hypothesized that retest effects on the job-knowledge test would be larger for White candidates than for Black and Hispanic candidates (Hypothesis 3), for females than for males (Hypothesis 4), and for younger candidates than for older candidates (Hypothesis 5). To test these hypotheses, we again used multiple regression analysis, whereby retest scores were regressed onto initial test scores and a dummy-coded variable that reflected candidates' standing on the demographic variable of interest. In this approach, the demographic variable is related to the residual of retest scores, which reflects the change from what is predicted based on initial test scores. If the beta coefficient for the demographic variable is significant, this indicates that demographics predict score change from the initial test to the retest. This approach also accounts for regression to the mean because change is measured as a deviation from the predicted value (Cohen & Cohen, 1983).

The regression results are shown in Table 4, and the associated initial–retest *d* statistics for each subgroup can be found in Table 3.

Table 4
Results of Multiple Regression Analyses Testing Subgroup Differences in Retest Effects

Subgroup comparison/predictor	*N*	*B*	*SE*	β
White–Black	85			
Initial test scores		0.62	0.14	.45*
Subgroup variable		1.57	1.26	.12
White–Hispanic	69			
Initial test scores		0.59	0.14	.46*
Subgroup variable		1.39	1.43	.11
White–Asian	139			
Initial test scores		0.65	0.10	.49*
Subgroup variable		−0.00	1.07	.00
Male–female	192			
Initial test scores		0.67	0.09	.50*
Subgroup variable		−2.18	0.95	−.15*
Age	190			
Initial test scores		0.65	0.08	.48*
Subgroup variable		−0.16	0.04	.23*

Note. Subgroup variables for race and gender were coded such that positive regression coefficients reflect larger score improvement for majority-group candidates (i.e., White and male candidates), whereas age was treated as continuous. *B* = unstandardized regression coefficient; *SE* = standard error of unstandardized regression coefficient; β = standardized regression coefficient.
* $p < .05$, two-tailed.

Although there was a trend for White candidates to improve more upon retesting than Black candidates ($d = 1.03$ vs. 0.81), the beta weight for the subgroup variable was nonsignificant. The regression comparing White versus Hispanic candidates also was nonsignificant, as the retest effects for these two groups were nearly identical ($d = 1.03$ vs. 1.02). Thus, Hypothesis 3 was not supported. Although we did not make predictions regarding possible retest effect differences between Asians and the other subgroups, the overall level of score improvement among Asian candidates ($d = 0.95$) was comparable to that of White and Hispanic candidates.

Table 4 also displays the regression results for gender and age, both of which were significant. Female candidates improved more than male candidates ($d = 1.23$ vs. 0.82), and younger candidates improved more than older candidates. Regarding age, candidates in their 20s and 30s improved by approximately the same amount ($d = 1.23$ and 1.13, respectively), whereas candidates in the 40s and 50s demonstrated notably less improvement upon retesting ($d = 0.75$ and 0.72, respectively). Finally, candidates in their 60s improved the least of all ($d = 0.41$). These results provide strong support for Hypotheses 4 and 5.

These subgroup differences in retest effects are important because they could influence subgroup differences in test scores, which have direct implications for adverse impact. To illustrate what effect retesting may have on subgroup differences, we also computed *d* statistics comparing scores of majority- and minority-group candidates on the initial test and on the retest. The results are provided in Table 5.

As shown, there were minimal changes in subgroup differences between White candidates and Black and Hispanic candidates from the initial test to the retest (although notice that the mean difference between Whites and Blacks became significant upon retesting). In contrast, the initial score difference that favored White over Asian candidates was reduced upon retesting. Subgroup differences for gender and age were consistent with the pattern of results in Tables 3 and 4. For gender, score differences on the initial test that favored male candidates were reversed upon retesting, such that females tended to receive higher test scores (although the mean difference was nonsignificant). For age, the initial (nonsignificant) score differences between younger and older candidates increased (and became significant) upon retesting to favor younger candidates. Overall, these results suggest that allowing candidates to retest can influence the magnitude and statistical significance of test-score differences between subgroups.

Effects of Retesting on Criterion-Related Validity

In Table 6, we present evidence concerning how retesting affects the criterion-related validity of job-knowledge test scores. Hypothesis 6 predicted that retest scores would be more predictive

[2] A reviewer suggested there may be a curvilinear relationship between time lag and retest effects. Specifically, perhaps there is an optimal time lag that is long enough to permit additional knowledge acquisition yet short enough to allow for memory of test content. We tested this possibility but found that the cross-product term for time lag did not predict retest scores beyond initial test scores and the linear time-lag variable, $\Delta R^2 = .00$, $F(1, 188) = 0.60$, $p = .44$.

Table 5
Subgroup Differences for Initial and Retest Promotion Test Scores

Subgroup comparison	*N*	Initial test *d*	Retest *d*
Whites vs. Blacks	85	0.42	0.44*
Whites vs. Hispanics	69	0.43	0.47
Whites vs. Asians	139	0.36*	0.17
Males vs. females	192	0.29	−0.18
Under 40 vs. 40 and older	190	0.08	0.45*

Note. d = standardized mean difference ($[M_{\text{majority group}} - M_{\text{minority group}}]/SD_{\text{pooled}}$).
*$p < .05$, two-tailed.

of job performance than would initial test scores. Job-performance ratings were available for 136 of the retest candidates. Within this group of retesters, validity estimates for initial and retest scores were .27 and .38, respectively. Correcting these estimates for criterion unreliability (using the .78 interrater estimate) yielded corrected validities of .31 and .43, respectively. A comparison of dependent correlations (Steiger, 1980) revealed that the observed validity estimate for retest scores was significantly larger than the observed validity estimate for initial scores, $t(133) = 1.66$, $p < .05$ (one-tailed).[3] These results provide support for Hypothesis 6 and suggest that among retesters, subsequent test scores were better predictors of job performance than were initial test scores.

A reviewer raised the interesting possibility that the criterion-related validity estimates for initial and retest scores may simply reflect a simplex pattern, whereby validity coefficients decrease in magnitude as the time lag between predictor and criterion measurement increases (Deadrick & Madigan, 1990; Henry & Hulin, 1989; Keil & Cortina, 2001). To investigate this possibility, we analyzed data from a subset of participants ($n = 52$) who failed the promotion test twice and then took it a third time. We did not consider these Time 3 data in our main analyses due to the small sample size, which, for example, would have made it impossible to examine subgroup differences in retesting. Support for a simplex pattern would be found if criterion-related validity increased from initial test to first retest to second retest, as test scores became more proximal to the time when job performance was measured. Within this group of 52 candidates, observed correlations between test scores and job performance were .32, .41, and .40, respectively, for the initial test, the first retest, and the second retest. Because validity appears to level off after the first retest, these results suggest that a simplex pattern of correlations may not account for the somewhat stronger validity evidence for retest scores.

Supplemental analyses. Although the above findings shed light on the relative validity of initial and retest scores, they do not directly address the types of practical retesting decisions that are made in organizations. Thus, we conducted some additional analyses to help illustrate how this finding may inform decisions about whether and how to offer retesting.

First, when deciding whether to initiate a retesting program, organizations may wish to know how allowing people to retest affects the overall criterion-related validity of a staffing procedure. Because we had criterion data on both successful and unsuccessful promotion candidates, we were able to investigate this issue. We first estimated the validity of the job-knowledge test using the initial test scores of all candidates who took the test, regardless of whether they passed or failed ($N = 403$). As Table 7 shows, the resulting observed validity estimate was .48. This reflects what the validity of the test would be in the absence of a retesting program.

We then estimated what the validity would be if the organization allowed retesting and the sample comprised both one-time and repeat candidates. We did this by replacing the initial test scores of candidates who failed the first time with their retest scores. Thus, we used the same set of candidates as before, but the predictor data now comprised a mix of initial scores (for candidates who passed the initial test) and retest scores (for the candidates who failed the initial test and chose to retake it). The resulting observed validity estimate was .51, which is slightly higher than the .48 estimate based on initial test scores only. This finding is consistent with the results pertaining to Hypothesis 6 and suggests that allowing candidates to retest would not negatively affect the overall validity of the promotion test.

Second, although we had job-performance information on both successful and unsuccessful candidates, organizations frequently have criterion data on successful candidates only. In situations such as this, one way to determine how implementing a retesting program may affect criterion-related validity is to examine whether passing scores of retesters predict performance on the job as well as passing scores of candidates who passed their initial test.

To illustrate, we estimated the validity for the 214 candidates in our sample who passed the promotion test on their first attempt and the validity for the 56 candidates who failed on their first attempt but passed on their second attempt. The resulting observed validity coefficients were .29 and .31, respectively (see Table 7; these *r*s are smaller than others we report due to the restricted range of scores from using data from passing candidates only). Moderated multiple regression can be used to test whether candidate status (one-time vs. repeat candidates) moderates criterion validity (Lievens et al., 2005). Given the similarity of validity coefficients from the two groups of candidates, it was not surprising that the interaction between candidate status and passing scores did not significantly increase the variance explained in job performance beyond the main effects of the two variables, $\Delta R^2 = .00$, $F(1, 266) = 0.08$, $p = .78$. This suggests that among candidates who eventually passed the promotion test, scores of retest candidates were as valid as scores of candidates who passed the initial test.

Finally, once organizations decide to implement a retesting program, they must determine which score(s) to use as a basis for selection or promotion. We examined four possible options: initial test scores, most recent test scores, highest test scores, and the mean of candidates' initial and retest scores. The corresponding criterion-related validities can be found in Table 7. The three options that incorporate candidates' retest scores were similarly predictive of performance, with mean test scores having the highest validity estimate ($r = .39$). All these options would provide a higher level of criterion-related validity than use of candidates' initial test scores only ($r = .27$).

[3] As with the regression to the mean analyses described above, we used the range restriction-corrected correlation between initial test scores and retest scores for this analysis.

Table 6
Descriptive Statistics and Criterion-Related Validity Estimates for Retest Candidates

	Promotion test		Job performance			
Variable	*M*	*SD*	*M*	*SD*	*r*	r_c
Initial test scores	27.48	4.97	4.44	1.17	.27*	.31
Retest scores	33.12	6.22	4.44	1.17	.38*	.43

Note. $N = 136$. r = observed validity coefficient; r_c = validity coefficient corrected for unreliability in the job-performance criterion.
* $p < .05$, two-tailed.

Discussion

The practice of retesting continues to attract the attention of both practitioners and researchers. This attention is understandable given that decisions about retesting may have substantial consequences for both organizations and job candidates. The purpose of the present study was to extend prior research by examining retest effects on a job-knowledge test used to facilitate promotion decisions. We also investigated the effects of retesting on two critical factors in the use of staffing procedures, namely, subgroup differences and criterion-related validity for predicting job performance. Furthermore, the within-job promotion design of the study allowed us to examine these factors in the absence of the detrimental influences of direct range restriction.

Key Findings and Implications

Overall retest effects. Previous research primarily has examined retest effects on cognitive ability or personality tests. Our examination of a test of job knowledge responds to calls for additional research to help understand retest effects on knowledge-oriented measures (e.g., Lievens et al., 2007). Approximately 70% of the candidates in our study who initially failed the promotion test chose to retest. This high percentage of retesters appears consistent with the results of a recent study by Hausknecht (2010), who found that internal candidates were over four times more likely to retest than external candidates. Hausknecht speculated that because internal candidates are closer to the internal job market and have lower job-search costs, they are more likely to retest than external applicants, who are farther removed from the organization and have more limited access to information about job openings, staffing procedures, and feedback concerning their initial test performance.

Candidates who chose to retest improved their scores upon retesting by almost a full standard deviation ($d = 0.93$), although almost one half of this improvement appears to be due to regression to the mean. The magnitude of this improvement was much larger than that reported by Schleicher et al. (2010; $d = 0.15$) but is more in line with that reported by Raymond et al. (2007) on two certification tests ($d = 0.79$ and 0.48). As with this latter study, participants in the present study remained in the job despite failing the initial test (in contrast, Schleicher et al., 2010, studied external applicants). This pattern of findings suggests that larger retest effects may be found in within-job promotion contexts in which candidates who initially fail can gain additional experience on the job, which may allow them to improve their test scores upon retesting. This possibility is consistent with the positive relationship we found between time lag and score improvement, whereby the longer candidates waited to retest, the more they improved (although the magnitude of this relationship was modest).

Subgroup differences in retest effects. This is one of the first studies to investigate the important issue of subgroup differences in retest effects. Our results reinforce those reported by

Table 7
Criterion-Related Validity Estimates From Supplementary Analyses

Question and comparison	*N*	*r*	r_c
How does allowing candidates to retest affect the overall criterion-related validity of job-knowledge test scores?			
Initial test scores only	403	.48	.54
Initial test scores and retest scores[a]	403	.51	.58
In the absence of criterion data on all candidates, are passing scores of repeat candidates more or less valid than passing scores of one-time candidates?			
Passing scores of one-time candidates	214	.29	.33
Passing scores of repeat candidates	56	.31	.35
For candidates who retest, which score is most predictive of job performance?			
Initial test score	136	.27	.31
Most recent test score	136	.38	.43
Highest score	136	.37	.42
Average of all scores	136	.39	.44

Note. All validity estimates are statistically significant ($p < .05$, two-tailed).
[a] Includes initial test scores for one-time applicants (who passed the initial test) and retest scores for repeat applicants (who failed the initial test and then retested).

Schleicher and colleagues (2010) in that females and younger candidates were more likely to improve than males and older candidates. The relatively larger improvements for female candidates provide support for the theoretical explanations we have offered, such as the fact that women tend to respond more favorably to negative feedback and make better use of that feedback than do men.

Our findings with regard to age extend those of Schleicher and colleagues (2010) by showing that score improvements appear to begin to decline around age 40 and then slowly disappear as candidates near retirement age. This pattern of results is consistent with research on fluid intelligence, which starts to decline in people's late 30s (Horn & Cattell, 1967). These results also are in line with theory that would predict that younger candidates, who are more likely to focus on growth and performance optimization, may be more motivated to improve upon retesting than older candidates, who are more likely to be concerned with avoiding losses (e.g., Baltes & Baltes, 1990; Freund & Baltes, 2000).

In addition, there is some evidence to suggest that Black candidates might not improve as much with retesting as candidates from other ethnic groups, which also is consistent with the findings of Schleicher and colleagues (2010). Nonetheless, this difference was not statistically significant. Several reasons may have contributed to this finding. For instance, the number of Black candidates in our sample was somewhat modest ($n = 34$), which limited the statistical power to detect significant differences in retest effects between this and other subgroups. In addition, candidates overall were highly educated, so the range of mental ability was somewhat restricted. This is relevant because subgroup differences in ability are one potential reason for ethnic group differences in retest effects.

From a practical perspective, the subgroup differences in retest effects we found are important because they may influence subgroup differences in test scores, which have direct implications for adverse impact. For example, we found that initial score differences that favored White candidates over Asian candidates and male candidates over female candidates were reduced, or even reversed, upon retesting. In contrast, initial score differences that favored younger candidates over older candidates increased upon retesting.

These results demonstrate how allowing candidates to retest can influence the magnitude and statistical significance of test-score differences between subgroups because candidates from some groups tended to improve more upon retesting than did candidates from other groups. Thus far, the results of the present study and those of Schleicher et al. (2010) converge to suggest that allowing candidates to retest may reduce adverse impact against female candidates but increase adverse impact against older candidates and possibly candidates of certain ethnic groups.

Effects of retesting on criterion-related validity. The present study provides some of the first data regarding the influence of retesting on criterion-related validity with respect to job performance. We discovered that among candidates who retested, retest scores were somewhat better predictors of job performance than were initial test scores. This finding is consistent with the limited research on retesting and validity in academic settings, which has found that retest scores tend to be better predictors of criteria (e.g., GPA) than initial test scores (e.g., Allalouf & Ben-Shakhar, 1998; Coyle, 2006; Lievens et al., 2005).

The finding of higher criterion-related validity for retest scores has implications for both theory and practice. From a theoretical perspective, this finding helps test and refine theories on retesting and validity. Specifically, our results do not appear to support the proposition that initial and retest scores capture the underlying construct to the same extent (Explanation 1 in Lievens et al., 2007) or the proposition that retesting increases the measurement of criterion-irrelevant factors, such as test-taking strategies (Explanation 3 in Lievens et al., 2007).

Instead, our findings appear to support the proposition that retesting either reduces construct contamination (Explanation 2 in Lievens et al., 2007) or increases variance in the target construct. For one, retest scores were significantly higher than initial test scores. This finding is consistent with the idea that retesting minimizes the influence of construct-irrelevant factors, such as test anxiety. Furthermore, as in past research (e.g., Schleicher et al., 2010), the variance in retest scores was larger than the variance in initial test scores. This occurred despite the fact that retest scores overall increased upon retesting, which would tend to limit variability.

The wider variance in job-knowledge scores upon retesting may be a result of individual differences in factors such as ability and motivation. For example, results from the subgroup analyses support the idea that females and younger candidates may be more motivated to pass the retest than males and older candidates. This, in turn, may lead to individual differences in test preparation and, ultimately, larger variance in job knowledge upon retesting. The role of motivation and test preparation also would be consistent with the positive relationship we found between score improvement and time between test occasions. Furthermore, the possibility that motivational factors influence retest scores is in line with Hausknecht et al. (2002), whose results suggested that candidates who chose to retake selection tests after failing possessed higher levels of motivation and persistence.

From a practical standpoint, our results suggest that allowing candidates to retest will not reduce the criterion-related validity of staffing procedures and may even increase it. Thus, not only may retesting expand applicant pools and improve corporate reputation, it also may lead to better selection decisions. The present results do not point to a preferred policy in terms of test-score use. Indeed, we observed similar levels of criterion-related validity when using candidates' most recent test scores, highest test scores, and average test scores. However, all three sets of scores provided better prediction of job performance than did candidates' initial test scores.

Boundary Conditions, Limitations, and Directions for Future Research

We conclude by noting some boundary conditions and limitations of this study, as well as some possible directions for future research on retesting. In terms of boundary conditions, it is important to remember that our results are based on data from internal candidates who completed a job-knowledge test for a within-job promotion and who retested an average of about 4 months after the initial test. Thus, our results are probably most relevant to retesting in situations involving knowledge-oriented measures completed by internal candidates. They also may have relevance for licensing tests used to credential members of various

occupations (e.g., accountants, clinicians, lawyers, physicians) and for certification tests taken by members of professional organizations (e.g., the Society for Human Resources Management). Although retesting in these contexts is widespread, there is minimal research to guide test developers and administrators (for an exception, see Raymond et al., 2007).

On the other hand, our results may be less likely to generalize to other types of retesting situations. For example, the magnitude and effects of retesting may be somewhat different in situations that involve longer retesting intervals (e.g., in which applicants must wait 12 months before reapplying), different types of selection procedures (e.g., those that measure more stable constructs, such as cognitive ability, or those that are susceptible to response distortion, such as personality measures), external applicants (e.g., who may be less likely to retest and who may receive more limited feedback about their initial test performance), or individuals who apply for positions other than a within-job promotion (e.g., between-job promotions in which candidates cannot benefit from additional experience on the job).

There also are some factors that limit the inferences that can be drawn on the basis of our results. First, although we were able to examine the nature and effects of retesting using a predictive design and employed individuals who were applying for a real promotion, this operational context limited our ability to collect data on variables that may have enabled us to test some of our theoretical explanations more directly. For instance, although several findings were consistent with the idea that retest scores may be somewhat better indicators of individual differences in job knowledge, we were unable to test this possibility directly. As an example, factors such as candidate demographic characteristics and the time lag between initial and retest likely are proxies for underlying mechanisms that ultimately drive retest effects, such as ability, motivation, attributions, and learning. A critical need for future research is to incorporate more direct measures of these and other possible underlying mechanisms to shed light on the "why" of retest effects. Additional research on the role of candidate demographics (e.g., in the propensity to retest) and process variables (e.g., length of waiting period, whether and how to combine initial and retest scores for use in decision making) in retesting also is important from both a legal and a workforce effectiveness standpoint.

Second, although our results are based upon a diverse sample of candidates with regard to race, gender, and age, the sample sizes for certain subgroups (namely, Blacks and Hispanics) were smaller than ideal. This may have limited our ability to detect statistically significant retest effects between these and other subgroups. Given that subgroup differences in retest effects may influence adverse impact, we call upon researchers to continue this important line of inquiry.

Third, we investigated how scores on an initial promotion test and one retest—taken an average of 4 months apart—relate to subsequent performance on the job. Future studies might attempt to incorporate multiple retests and more varied retesting time intervals. Future research also might consider whether and how retesting affects other criteria. For example, people who choose to retest may be particularly committed to a given organization. If so, applicants who succeed upon retesting may be less likely to leave the organization.

Finally, very little is known about what types of candidates choose to retest in the first place. In one of the first studies to start to address this issue, Hausknecht (2010) reported that internal candidates were more likely to retest than were external candidates. As an ancillary analysis, we explored whether candidate subgroup (e.g., males vs. females) influenced decisions to retest in our sample, but it did not. However, it appears that candidates who scored higher on the initial promotion test tended to retest sooner than candidates who scored lower on the initial test ($r = -.22$; see Table 2). An important direction for future research is to identify other variables that may affect decisions to retest.

References

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. doi:10.1111/j.1745-3984.1998.tb00526.x

Alliger, G. M., & Katzman, S. (1997). When training affects variability: Beyond the assessment of mean differences in training evaluation. In J. K. Ford (Ed.), *Improving training effectiveness in work organizations* (pp. 223–243). Mahwah, NJ: Erlbaum.

Arvey, R. D., Strickland, W., Drauden, G., & Martin, C. (1990). Motivational components of test taking. *Personnel Psychology, 43,* 695–716. doi:10.1111/j.1744-6570.1990.tb00679.x

Baltes, P. B., & Baltes, M. M. (1990). Psychological perspectives on successful aging: The model of selective optimization with compensation. In P. B. Baltes & M. M. Baltes (Eds.), *Successful aging: Perspectives from the behavioral sciences* (pp. 1–34). New York, NY: Cambridge University Press. doi:10.1017/CBO9780511665684.003

Bauer, T. N., Maertz, C. P., Jr., Dolen, M. R., & Campion, M. A. (1998). Longitudinal assessment of applicant reactions to employment testing and test outcome feedback. *Journal of Applied Psychology, 83,* 892–903. doi:10.1037/0021-9010.83.6.892

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

Boggiano, A. K., & Barrett, M. (1991). Gender differences in depression in college students. *Sex Roles, 25,* 595–605. doi:10.1007/BF00289566

Bourdeau, N. R. (2008, April). *Applicant retesting policy: Key considerations and best practices.* Panel discussion presented at the 23rd Annual Conference of the Society for Industrial and Organizational Psychology, San Francisco, CA.

Cassady, J. C., & Johnson, R. E. (2002). Cognitive test anxiety and academic performance. *Contemporary Educational Psychology, 27,* 270–295. doi:10.1006/ceps.2001.1094

Cattell, R. B. (1987). *Intelligence: Its structure, growth, and action.* Amsterdam, the Netherlands: North-Holland.

Chan, D. (1997). Racial subgroup differences in predictive validity perceptions on personality and cognitive ability tests. *Journal of Applied Psychology, 82,* 311–320. doi:10.1037/0021-9010.82.2.311

Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology, 82,* 143–159. doi:10.1037/0021-9010.82.1.143

Chan, D., Schmitt, N., DeShon, R. P., Clause, C. S., & Delbridge, K. (1997). Reactions to cognitive ability tests: The relationships between race, test performance, face validity perceptions, and test-taking motivation. *Journal of Applied Psychology, 82,* 300–310. doi:10.1037/0021-9010.82.2.300

Chan, D., Schmitt, N., Sacco, J. M., & DeShon, R. P. (1998). Understanding pretest and posttest reactions to cognitive ability and personality tests. *Journal of Applied Psychology, 83,* 471–485. doi:10.1037/0021-9010.83.3.471

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

Coyle, T. R. (2006). Test-retest changes on scholastic aptitude tests are not related to g. *Intelligence, 34,* 15–27. doi:10.1016/j.intell.2005.04.001

Deadrick, D. L., & Madigan, R. M. (1990). Dynamic criteria revisited: A longitudinal study of performance stability and predictive validity. *Personnel Psychology, 43,* 717–744. doi:10.1111/j.1744-6570.1990.tb00680.x

Duehr, E. E., Jackson, H. L., & Ones, D. S. (2004, April). *Gender differences in Big Five factors and facets: A meta-analysis.* Paper presented at the 19th Annual Conference of the Society for Industrial and Organizational Psychologists, Chicago, IL.

Ebner, N. C., Freund, A. M., & Baltes, P. B. (2006). Developmental changes in personal goal orientation from young to late adulthood: From striving for gains to maintenance and prevention of losses. *Psychology and Aging, 21,* 664–678. doi:10.1037/0882-7974.21.4.664

Ellingson, J. E., Sackett, P. R., & Connelly, B. S. (2007). Personality assessment across selection and development contexts: Insights into response distortion. *Journal of Applied Psychology, 92,* 386–395. doi:10.1037/0021-9010.92.2.386

Feingold, A. (1994). Gender differences in personality: A meta-analysis. *Psychological Bulletin, 116,* 429–456. doi:10.1037/0033-2909.116.3.429

Fisher, J. D., Harrison, C. L., & Nadler, A. (1978). Exploring generalizability of donor-recipient similarity effects. *Personality and Social Psychology Bulletin, 4,* 627–630. doi:10.1177/014616727800400428

Freund, A. M. (2006). Age-differential motivational consequences of optimization versus compensation focus in younger and older adults. *Psychology and Aging, 21,* 240–252. doi:10.1037/0882-7974.21.2.240

Freund, A. M., & Baltes, P. B. (2000). The orchestration of selection, optimization and compensation: An action-theoretical conceptualization of a theory of developmental regulation. In W. J. Perrig & A. Grob (Eds.), *Control of human behavior, mental processes, and consciousness: Essays in honor of the 60th birthday of August Flammer* (pp. 35–58). Mahwah, NJ: Erlbaum.

Gilliland, S. W. (1993). The perceived fairness of selection systems: An organizational justice perspective. *Academy of Management Review, 18,* 694–734. doi:10.2307/258595

Greene, K. (2005, September 26). When we're all 64. *The Wall Street Journal,* (pp. R1–R4).

Hausknecht, J. P. (2010). Candidate persistence and personality test practice effects: Implications for staffing system management. *Personnel Psychology, 63,* 299–324. doi:10.1111/j.1744-6570.2010.01171.x

Hausknecht, J. P., Day, D. V., & Thomas, S. C. (2004). Applicant reactions to selection procedures: An updated model and meta-analysis. *Personnel Psychology, 57,* 639–683. doi:10.1111/j.1744-6570.2004.00003.x

Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., & Moriarty Gerrard, M. O. (2007). Retesting in selection: A meta-analysis of practice effects for tests of cognitive ability. *Journal of Applied Psychology, 92,* 373–385. doi:10.1037/0021-9010.92.2.373

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. doi:10.1037/0021-9010.87.2.243

Heckhausen, J. (1997). Developmental regulation across adulthood: Primary and secondary control of age-related challenges. *Developmental Psychology, 33,* 176–187. doi:10.1037/0012-1649.33.1.176

Helms, J. E. (1992). Why is there no study of cultural equivalence in standardized cognitive ability testing? *American Psychologist, 47,* 1083–1101. doi:10.1037/0003-066X.47.9.1083

Henry, R. A., & Hulin, C. L. (1989). Changing validities: Ability-performance relations and utilities. *Journal of Applied Psychology, 74,* 365–367. doi:10.1037/0021-9010.74.2.365

Hirschy, A. J., & Morris, J. R. (2002). Individual differences in attributional style: The relational influence of role self-efficacy, self-esteem, and sex role identity. *Personality and Individual Differences, 32,* 183–196. doi:10.1016/S0191-8869(01)00003-4

Hogan, J., Barrett, P., & Hogan, R. (2007). Personality measurement, faking, and employment selection. *Journal of Applied Psychology, 92,* 1270–1285. doi:10.1037/0021-9010.92.5.1270

Horn, J. L., & Cattell, R. B. (1967). Age differences in fluid and crystallized intelligence. *Acta Psychologica, 26,* 107–129. doi:10.1016/0001-6918(67)90011-X

Hunter, J. E., Schmidt, F. L., & Le, H. (2006). Implications for direct and indirect range restriction for meta-analysis methods and findings. *Journal of Applied Psychology, 91,* 594–612. doi:10.1037/0021-9010.91.3.594

Hyvönen, K., Feldt, T., Salmela-Aro, K., Kinnunen, U., & Makikangas, A. (2009). Young managers' drive to thrive: A personal work goal approach to burnout and work engagement. *Journal of Vocational Behavior, 75,* 183–196. doi:10.1016/j.jvb.2009.04.002

Jensen, A. R. (1998). *The g factor: The science of mental ability.* London, England: Praeger.

Johnson, M., & Helgeson, V. S. (2002). Sex differences in response to evaluative feedback: A field study. *Psychology of Women Quarterly, 26,* 242–251. doi:10.1111/1471-6402.00063

Kanfer, R., & Ackerman, P. L. (2004). Aging, adult development, and work motivation. *Academy of Management Review, 29,* 440–458. doi:10.2307/20159053

Keil, C. T., & Cortina, J. M. (2001). Degradation of validity over time: A test and extension of Ackerman's model. *Psychological Bulletin, 127,* 673–697. doi:10.1037/0033-2909.127.5.673

Kelley, P. L., Jacobs, R. R., & Farr, J. L. (1994). Effects of multiple administrations of the MMPI for employee screening. *Personnel Psychology, 47,* 575–591. doi:10.1111/j.1744-6570.1994.tb01738.x

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.

Lievens, F., Buyse, T., & Sackett, P. R. (2005). Retesting effects in operational selection settings: Development and test of a framework. *Personnel Psychology, 58,* 981–1007. doi:10.1111/j.1744-6570.2005.00713.x

Lievens, F., Reeve, C. L., & Heggestad, E. D. (2007). An examination of psychometric bias due to retesting on cognitive ability tests in selection settings. *Journal of Applied Psychology, 92,* 1672–1682. doi:10.1037/0021-9010.92.6.1672

Macoby, E., & Jacklin, C. (1974). *The psychology of sex differences.* Stanford, CA: Stanford University Press.

McCarthy, J. M., & Goffin, R. D. (2005). Selection test anxiety: Exploring tension and fear of failure across the sexes in simulated selection scenarios. *International Journal of Selection and Assessment, 13,* 282–295. doi:10.1111/j.1468-2389.2005.00325.x

McGehee, W., & Thayer, P. W. (1961). *Training in business and industry.* New York, NY: Wiley.

Millman, J., Bishop, H., & Ebel, R. (1965). An analysis of test-wiseness. *Educational and Psychological Measurement, 25,* 707–726. doi:10.1177/001316446502500304

Rapport, L. J., Brines, D. B., Axelrod, B. N., & Theisen, M. E. (1997). Full scale IQ as mediator of practice effects: The rich get richer. *Clinical Neuropsychologist, 11,* 375–380. doi:10.1080/13854049708400466

Raymond, M. R., Neustel, S., & Anderson, D. (2007). Retest effects on identical and parallel forms in certification and licensure testing. *Personnel Psychology, 60,* 367–396. doi:10.1111/j.1744-6570.2007.00077.x

Reeve, C. L., & Lam, H. (2005). The psychometric paradox of practice effects due to retesting: Measurement invariance and stable ability estimates in the face of observed score changes. *Intelligence, 33,* 535–549. doi:10.1016/j.intell.2005.05.003

Reeve, C. L., & Lam, H. (2007). The relation between practice effects, test-taker characteristics and degree of *g*-saturation. *International Journal of Testing, 7,* 225–242. doi:10.1080/15305050701193595

Roberts, T. A., & Nolen-Hoeksema, S. (1989). Sex differences in reaction to evaluative feedback. *Sex Roles, 21,* 725–747. doi:10.1007/BF00289805

Roberts, T. A., & Nolen-Hoeksema, S. (1994). Gender comparisons in responsiveness to others' evaluations in achievement settings. *Psychology of Women Quarterly, 18,* 221–240. doi:10.1111/j.1471-6402.1994.tb00452.x

Roediger, H. L., III, & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1,* 181–210. doi:10.1111/j.1745-6916.2006.00012.x

Roth, P. L., BeVier, C. A., Bobko, P., Switzer, F. S., III, & Tyler, P. (2001). Racial differences in cognitive abilities: A meta-analysis. *Personnel Psychology, 54,* 297–330. doi:10.1111/j.1744-6570.2001.tb00094.x

Roth, P. L., Van Iddekinge, C. H., Huffcutt, A. I., Eidson, C. E., Jr., & Bobko, P. (2002). Correcting for range restriction in structured interview ethnic group differences: The values may be larger than researchers thought. *Journal of Applied Psychology, 87,* 369–376. doi:10.1037/0021-9010.87.2.369

Sachau, D. A., Houlihan, D., & Gilbertson, T. (1999). Predictors of employee resistance to supervisors' requests. *Journal of Social Psychology, 139,* 611–621. doi:10.1080/00224549909598421

Schaie, K. W. (1996). *Intellectual development in adulthood: The Seattle longitudinal study.* New York, NY: Cambridge University Press.

Schleicher, D. J., Van Iddekinge, C. H., Morgeson, F. P., & Campion, M. A. (2010). If at first you don't succeed, try, try again: Understanding race, gender, and age differences in retesting test score improvement. *Journal of Applied Psychology, 95,* 603–617. doi:10.1037/a0018920

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel selection: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124,* 262–274. doi:10.1037/0033-2909.124.2.262

Schmidt, F. L., Hunter, J. E., & Outerbridge, A. N. (1986). Impact of job experience and ability on job knowledge, work sample performance, and supervisory ratings of job performance. *Journal of Applied Psychology, 71,* 432–439. doi:10.1037/0021-9010.71.3.432

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87,* 245–251. doi:10.1037/0033-2909.87.2.245

te Nijenhuis, J., van Vianen, A. E. M., & van der Flier, H. (2007). Score gains on *g*-loaded tests: No *g. Intelligence, 35,* 283–300.

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. doi:10.1111/1468-2389.00182

Thorndike, R. L. (1949). *Personnel selection.* New York, NY: Wiley.

U.S. Bureau of Labor Statistics. (2009). *Women in the labor force: A databook* (Report No. 1018). Washington, DC: U.S. Department of Labor.

Vernon, P. E. (1954). Practice and coaching effects in intelligence tests. *Educational Forum, 18,* 269–280. doi:10.1080/00131725409341273

Wheeler, J. K. (2004, April). *Practical implications of selection retesters on testing development and policy.* Practitioner forum presented at the 19th Annual Conference of the Society for Industrial and Organizational Psychology, Chicago, IL.

Received November 19, 2009
Revision received December 15, 2010
Accepted January 17, 2011 ■