Journal of Applied Psychology
2010, Vol. 95, No. 4, 603–617

0021-9010/10/$12.00 DOI: 10.1037/a0018920

If At First You Don't Succeed, Try, Try Again: Understanding Race, Age, and Gender Differences in Retesting Score Improvement

Deidra J. Schleicher
Purdue University

Chad H. Van Iddekinge
Florida State University

Frederick P. Morgeson
Michigan State University

Michael A. Campion
Purdue University

This article explores the intersection of 2 critical and timely concerns in personnel selection—applicant retesting and subgroup differences—by exploring demographic differences in retest effects across multiple assessments. Results from large samples of applicants taking 3 written tests ($N = 7{,}031$) and 5 performance tests ($N = 2{,}060$) revealed that Whites showed larger retest score improvements than Blacks or Hispanics on several of the assessments. However, the differential improvement of Whites was greater on the written tests than on the performance tests. In addition, women and applicants under 40 years of age showed larger improvements with retesting than did men and applicants over 40. We offer some preliminary theoretical explanations for these demographic differences in retesting gains, including differences in ability, testing attitudes and motivation, and receptivity to feedback. In terms of practical implications, the results suggest that allowing applicants to retake selection tests may, in some cases, exacerbate levels of adverse impact, which can have distinct implications for retesting policy and practices in organizations.

Keywords: personnel selection, employment tests, retesting, practice effects, demographic differences

The outcome of employee selection processes is a consequence of considerable importance to applicants and organizations alike. Not only do applicants wish to obtain an offer of employment, they also desire to be given every opportunity to perform during the selection process so they can demonstrate their qualifications for the job (Schleicher, Venkataramani, Morgeson, & Campion, 2006). For their part, organizations want to ensure that they have the best information available on candidates so they can make informed selection decisions. Yet, for a variety of reasons, applicants do not always perform at their best during the selection process, resulting in potential "false negative" decisions. Thus, it is not surprising that policies that allow for retesting have been encouraged by both legal and professional guidelines on employee selection (e.g., Society for Industrial and Organizational Psychology, 2003; Uniform Guidelines on Employee Selection Procedures, 1978).

Concerns about corporate reputation have also likely motivated organizations to offer retesting opportunities to initially unsuccessful applicants. Indeed, the opportunity for reconsideration is a key component of procedural justice perceptions (Arvey & Sackett, 1993; Gilliland, 1993; Gilliland & Steiner, 2001). Justice perceptions, in turn, have been linked to applicants' evaluations of the organization, intentions to accept job offers, and propensity to recommend the employer to others (Hausknecht, Day, & Thomas, 2004). For these reasons, many organizations in the private and public sectors have implemented retesting policies in hiring and promotion situations, and a large number of applicants take advantage of these opportunities (Hausknecht, Halpert, Di Paolo, & Moriarty Gerrard, 2007; Lievens, Buyse, & Sackett, 2005; Wheeler, 2004).

The prevalence of applicant retesting has stimulated research that attempts to better understand the nature and implications of this practice. This research has reached two clear conclusions. First, there are consistent score improvements due to retesting on cognitive ability tests (Hausknecht et al., 2007; Hausknecht, Trevor, & Farr, 2002; Kulik, Kulik, & Bangert, 1984; Lievens et al., 2005; Reeve & Lam, 2005, 2007; Sackett, Burris, & Ryan, 1989). Second, these increases can have considerable implications for exactly who is hired. As Hausknecht et al. (2002) noted, "systematic score changes across repeated administrations can result in a qualitative change in the make-up of the workforce" (p. 244). For example, in a recent meta-analysis on cognitive ability tests (Hausknecht et al., 2007), the average size of retesting effects (corrected for sampling error and measurement error) indicated that applicants in the 50th percentile at Time 1 would move to the 60th percentile at Time 2 and to the 71st percentile at Time 3. In

Deidra J. Schleicher and Michael A. Campion, Krannert School of Management, Purdue University; 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 Chad H. Van Iddekinge contributed equally to this study.

Correspondence concerning this article should be addressed to Deidra J. Schleicher, Purdue University, Rawls Hall Room 4034, 100 South Grant, West Lafayette, IN 47907-2076. E-mail: deidra@purdue.edu

other words, applicants substantially below the score cutoff in the first testing could eventually score well enough to be hired.

Although this research has increased the understanding of various issues associated with applicant retesting, other important issues have received little or no research attention. First, past research has not investigated whether different groups of applicants show differential levels of score improvement. In particular, relatively little is known about the extent to which applicants of different racial, gender, and age groups demonstrate different levels of change across repeated test administrations. This is a critical issue to examine, because such differences could have substantial implications for adverse impact. That is, all other things being equal, if score increases due to retesting are greater for members of typically disadvantaged groups, adverse impact could be less at retesting than at initial testing. However, if score increases due to retesting are smaller for typically disadvantaged groups, levels of adverse impact could be exacerbated at retesting. A third possibility is that score changes are relatively similar across demographic groups (i.e., everyone's scores change by essentially the same amount). In this case, assuming that no demographic differences exist in decisions to retest (e.g., Schmit & Ryan, 1997), there should be little or no difference in adverse impact between the initial test and retest. Thus, this question of demographic differences in test score improvement with retesting is one with significant practical implications for organizations. It also has important theoretical implications for the understanding of what underlies retesting effects.

Second, previous research on retesting effects has focused predominately on cognitive ability tests (Hausknecht et al., 2002, 2007; Lievens, Reeve, & Heggestad, 2007; Reeve & Lam, 2005, 2007) and personality tests (Ellingson, Sackett, & Connelly, 2007; Hogan, Barrett, & Hogan, 2007; Kelley, Jacobs, & Farr, 1994). Yet there are many other types of assessments used by organizations (e.g., interviews, biographical data measures, assessment center exercises) that might be retaken by applicants, but for which virtually nothing is known about the nature or degree of retest effects. Thus, it is unknown whether the retest effects found for cognitive ability and personality tests generalize to other widely used assessments and whether any demographic differences in retesting effects vary across type of assessment.

Third, the majority of past retesting research has been conducted with student populations or in laboratory settings. For example, the vast majority of retesting studies included in a recent meta-analysis (Hausknecht et al., 2007) were conducted in an educational (as opposed to work) context (84%; J. Hausknecht, personal communication, 2008). Although studies conducted in educational settings have made important contributions to our understanding of retest effects, it is unclear whether similar effects would be found in employment settings. Therefore, there is a need for retesting research conducted in authentic job applicant contexts.

Finally, existing retesting research is characterized by relatively small samples. For example, Hausknecht et al. (2007) reported a median *N* of 91 participants per sample in their meta-analysis. One of the challenges associated with smaller samples (in addition to lower power) is the potentially restricted range of different kinds of test takers, including representation of applicants from certain demographic subgroups.

In the current article, we examine the practically and theoretically important question of whether race, age, and gender differences exist in retest effects. We do so across eight selection tests: three written tests (i.e., verbal ability, job knowledge, and biodata), three types of interviews (i.e., behavior description, situational, and an experience and interest interview), and two assessment center exercises (i.e., a leaderless group discussion and a case analysis exercise). To our knowledge, several of these assessments have not been previously studied for retest effects. Furthermore, we explore these effects across large and demographically diverse samples of actual job applicants ($N = 7{,}031$ for the written tests; $N = 2{,}060$ for the performance tests).

Hypothesis Development

A number of possible causes of retesting effects have been offered in the literature, some of which should vary between demographic groups and therefore inform our hypotheses. Lievens et al. (2005) delineated four main factors that could affect retest effects. First, such effects could reflect change due to random measurement error, which could result in higher or lower observed scores upon retesting. Second, retest effects could reflect true changes in the construct of interest, which also could result in higher or lower test scores on retesting (although the typical assumption is improvement with retesting; Lievens et al., 2005; Raymond, Neustel, & Anderson, 2007). Third, retest effects could reflect a criterion-relevant change in the observed score but no true change in the construct of interest. This has the effect of reducing or eliminating a deficit between the observed score and the true score, making the observed score a better approximation of the true score. Examples of this sort of change include a reduction in test unfamiliarity or anxiety and/or an increase in test motivation or self-efficacy (Hausknecht, Halpert, Harder, Kuljanin, & Moriarty, 2005). Finally, retest effects could reflect a criterion-irrelevant change in the observed score (e.g., learning "tricks" or other forms of artificial improvement) but no true change in the construct of interest.

Some of these proposed factors are likely to vary between demographic groups and therefore provide an a priori rationale for making predictions about such differences in score improvement. Most notably, there are theoretical arguments germane to how criterion-relevant changes in observed scores (e.g., due to anxiety or stress, test-taking motivation and self-efficacy), as well as true changes in the construct of interest, might vary across race. These are reviewed below, leading to our predictions for race-based differences in score improvements.

Race Differences in Score Improvements

If an examination of race differences in retesting effects revealed that minority (i.e., Black and Hispanic[1]) applicants show larger score gains with retesting, this could be good news for organizations concerned about adverse impact. However, available

[1] Our race-based hypotheses focus on Whites versus Blacks and Hispanics, because these groups represent the largest racial groups in the United States. In addition, Asian–White differences tend to be small and typically do not lead to adverse impact against applicants of Asian descent (Roth et al., 2001). However, to be comprehensive, we also report findings for Asian applicants, although we do not offer a priori hypotheses about these differences.

theory, as well as empirical evidence from other research areas, appears to support the opposite prediction: that White applicants will show greater score increases with retesting, particularly on the written tests. Two distinct arguments support this prediction, one involving differences in test attitudes and one involving ability-related score gains.

First, it has been suggested that testing attitudes, such as test self-efficacy, test-taking motivation, and perceived validity, vary between Black and White applicants (Chan, 1997; Chan & Schmitt, 1997; Ryan, 2001); White applicants generally hold more positive beliefs in these areas than minority applicants, particularly for cognitively oriented tests (Arvey, Strickland, Drauden, & Martin, 1990; Chan, Schmitt, Sacco, & DeShon, 1998). It has been further suggested that such differences may account for some of the racial differences in test scores (see Ryan, 2001, for a review). Thus, if "test attitudes can influence performance scores during each administration, it is conceivable that those with more favorable attitudes may also gain more from practice than those with negative attitudes" (Reeve & Lam, 2007, p. 229). In fact, Reeve and Lam (2007) found that test score gains generally were positively associated with undergraduate test-taker beliefs about the tests and their own motivation and ability to perform well on those tests. Thus, this line of research suggests that White applicants, because of their generally more favorable testing attitudes and motivation, will show larger score gains with retesting than minority applicants.

Second, research also suggests that score gains vary as a function of differences in examinee ability. In support of a "rich get richer" phenomenon (Rapport, Brines, Axelrod, & Theisen, 1997), Kulik et al. (1984) found retest effects (d, uncorrected) of 0.17, 0.40, and 0.80 for examinees of low, medium, and high ability, respectively. Higher ability candidates appear to improve more upon retesting than lower ability candidates, possibly because people with higher levels of cognitive ability tend to gain more from experience than do people with lower levels of ability (Jensen, 1998; Reeve & Lam, 2007). Because Whites tend to score higher on measures of cognitive ability than Blacks and Hispanics (Hough, Oswald, & Ployhart, 2001; Roth, BeVier, Bobko, Switzer, & Tyler, 2001; Ryan, 2001), the implication is that we may find larger score gains for White applicants. Thus, for both of the above reasons, we generally expect larger score gains with retesting for White applicants than for Black or Hispanic applicants.

> *Hypothesis 1:* There will be larger score gains with retesting for White applicants than for Black and Hispanic applicants.

These theoretical arguments and empirical findings about the role of test attitudes and ability also suggest a further expectation: that the differential retesting gains in favor of White applicants will be greater on the written tests than on the performance tests. This is because it is on written, standardized tests that the testing attitude differences between Whites and minorities tend to be most pronounced (Chan, 1997; Chan & Schmitt, 1997; Ryan, 2001). In addition, these are the types of assessments on which differences in cognitive ability might have the greatest impact. In fact, previous meta-analytic research has established that the closer a test gets to job performance (moving from written tests to performance-oriented tests, for example), the smaller the difference in scores between Whites and Blacks or Hispanics. For example, the White–Black and White–Hispanic d scores are 1.10 and 0.72 for cognitive ability tests (Roth et al., 2001), 0.55 and 0.67 for job knowledge tests (Roth, Huffcutt, & Bobko, 2003), 0.52 and 0.45 for work sample tests (Roth et al., 2003), and 0.27 and 0.04 for job performance ratings (Roth et al., 2003; see also McKay & McDaniel, 2006). These findings suggest that the differential ability advantage for White applicants may be more relevant for written tests (which are further from job performance) than for performance tests. Thus, we also predict the following:

> *Hypothesis 2:* The differential retesting gains in favor of White applicants will be greater on the written tests than on the performance tests.

Age and Gender Differences in Score Improvements

Although our primary focus is on racial differences in retest effects (because race represents the most common form of adverse impact in selection contexts; Hough et al., 2001), many of the same theoretical arguments noted above also apply to potential age and gender differences in score improvements. In addition, age and gender (like race) are protected classes for which there are sometimes score differences that could lead to differential selection (e.g., Saad & Sackett, 2002), thus suggesting the practical importance of these demographic variables as well. As such, we also examine age and gender differences in retest effects.

There is some evidence that both test-taking attitudes and motivation and ability-related differences may lead to differential retesting gains by age group. With respect to test-taking attitudes and motivation, it has been noted that older test-takers may have less positive attitudes toward testing in general and their own self-efficacy for doing well on tests, especially given that the bases of motivation tend to change across the lifespan (e.g., older workers tend to be less motivated by getting ahead and generative motives; Ebner, Freund, & Baltes, 2006; Freund, 2006; Kanfer & Ackerman, 2004). Ackerman, Beier, and Bowen (2002) noted that as individuals transition into middle age and beyond, they tend to see themselves as possessing lower ability for many tasks (especially those related to fluid intelligence, which is likely to be relevant to many selection tests; see below). As such, they tend to have a reduced interest in expending effort on such tasks (Kanfer & Ackerman, 2004). In addition, some research has found evidence of a small negative relationship between age and perceptions of procedural fairness on selection tests, such that older applicants tended to have somewhat lower perceptions of test fairness (see Hausknecht et al., 2004). As argued above, if such test attitudes and motivationally relevant self-perceptions can influence scores during each test administration, it is likely that these attitudes can also influence how much one gains between initial test and retest (Reeve & Lam, 2007).

There is also evidence to suggest that skills and abilities related to the likelihood of improving selection test performance, such as test-taking skills and fluid intelligence, decline with age. Sarnacki (1979, p. 264) noted that test-taking skills could be severely eroded by the passage of time, such that talents acquired through test-taking in high school and college could dissipate in older test-takers, especially those lacking recent exposure to testing. Fluid intelligence is the ability to think and reason abstractly and to solve problems; it is generally considered independent of learning, experience, and education (Cattell, 1987; Horn & Cattell, 1967). This type of intelligence is likely to be important for performing well and improving on selection tests, because many of these tests are designed to measure capabilities independent of specific learning and experiences and because test score improvement would require problem-solving about what went wrong on the initial test and how to improve on the retest. It is important to note that fluid intelligence peaks in adolescence and begins to decline progressively beginning in the 30s and 40s (Horn & Cattell, 1967; Kanfer & Ackerman, 2004; Schaie, 1996). Age-related declines in fluid intelligence, combined with the possibility of less positive testing attitudes and motivation in older test takers noted above, lead to an overall expectation that older applicants (i.e., those 40 and over) will demonstrate smaller retesting score gains than younger applicants.

> *Hypothesis 3:* There will be larger score gains for younger applicants (i.e., less than 40) than for older applicants (i.e., 40 and older).

Regarding gender, the expectations for differential test score improvements as a function of either ability-related or motivational and attitudinal differences are less clear. For example, there does not appear to be consistent evidence for any ability-related differences across the genders relevant to the selection tests used in the current context (e.g., we do not examine a test of quantitative ability, for which gender differences in ability or anxiety may be more relevant, see Nguyen & Ryan, 2008). With regard to testing attitudes, a recent meta-analysis examining applicant reactions to selection tests and testing procedures found no evidence of gender differences in such perceptions (see Hausknecht et al., 2004).

A few gender differences have been noted on some personality variables relevant to test motivation. For example, Hough et al. (2001) reported in their meta-analysis that women have somewhat higher scores on conscientiousness (particularly the dependability component of conscientiousness). However, these were relatively small effects (uncorrected ds $= -0.08$ and -0.22, respectively). In addition, in a large scale cross-cultural meta-analysis, Duehr, Jackson, and Ones (2004) concluded that any male–female differences on the Big Five personality traits were small and, when aggregated to the facet level, became negligible. Thus, because of the lack of a strong evidence base or theoretical arguments for expecting gender differences in either ability or testing attitudes and motivation, we do not offer a specific hypothesis with respect to gender differences in retesting score gains. Rather, we examine any gender differences in retest effects in an exploratory fashion.

Method

Participants and Procedure

Study participants were applying for professional positions with an agency of the U.S. government. These positions entailed working with the public, government officials, and the business community. Selected employees would work in one of several different career paths, including general management and specialty areas.

Approximately 15,000 persons apply for these positions in a given year. The first hurdle in the selection process includes a battery of three written tests (described below) designed to prescreen candidates on the basis of minimum qualifications. This set of tests is offered once a year. Approximately 40% of applicants achieve a qualifying score on a composite of the written tests and move on to the second stage of the process, which comprises three interviews and two assessment center exercises (described below). Approximately 25% pass the second hurdle composite and are placed on a "hirable" list from which job offers are made. (We subsequently refer to the first hurdle tests as the *written tests* and the second hurdle tests as the *performance tests*.) Applicants are allowed to retake the written tests and the performance tests. However, if they fail the performance tests (after passing the written tests), they must start at the beginning of the process and retake the written tests before retaking the performance tests.

The data for this study were collected from six test administrations. During this time, 7,031 applicants failed the written tests at Time 1 and retook the tests 1 year later. These individuals made up our written test sample. In addition, 2,060 applicants failed the performance tests at Time 1, returned 1 year later to retake the written tests, passed those tests, and then retook the performance tests. For control purposes, we only examined score changes between the initial test and the first retest. Further, to standardize time lag, we only included applicants who had failed the test the first time and who had taken the written retest 1 year after the initial test. (The performance tests are not offered on a standard schedule like the written tests are, but the average time-lag between initial performance test and the first retest was 1.3 years, which is very similar to the 1-year lag for the written tests.) Note that although the applicants in the performance test sample are also in the written test sample (i.e., they are a subset of the written test takers), we could not match applicants across the two samples because of incomplete identifying information. Thus, analyses are conducted separately on the two samples.

Table 1 provides demographic information for applicants included in these two samples. In addition, applicants across the two samples had a mean age of approximately 26 years; more than 90% had a bachelor's degree, and approximately 50% also had graduate or law degrees.

Description of Assessments[2]

Written tests. The three written tests included verbal ability, job knowledge, and biodata tests. These tests were administered on a single day in testing sessions that lasted a total of approximately 5 hr. The verbal ability test measured knowledge of correct grammar and the organization, word usage, spelling, and punctuation required for written reports and for editing the written work of others. The job knowledge test measured knowledge on a number of topics determined by a job analysis to be important for performing the tasks in this position. Despite the existence of multiple career paths, a single job knowledge test was used that assessed the common aspects of this position across the five career paths. The biodata test included items designed to measure job-related experience, skills, and achievements in school, work, and other areas.

[2] Given test security issues and the high-stakes nature of this context (in which applicants are extremely motivated to secure these positions), we describe these assessments only in broad terms. Researchers interested in more specific details (including the dimensions targeted in each assessment) are encouraged to contact the authors.

Table 1
Demographic Characteristics for Applicants in Written and Performance Tests Samples

	Written test sample		Performance test sample	
Characteristic	*N*	%	*N*	%
Race				
White	4,364	62.7	1,649	82.1
Black	898	12.9	65	3.3
Hispanic	793	11.4	125	6.2
Asian	904	13.0	171	8.5
Total	6,959		2,010	
Gender				
Male	4,143	58.9	1,336	65.0
Female	2,888	41.1	717	35.0
Total	7,031		2,053	
Age				
Under 40	5,785	82.3	1,740	85.7
40 and older	1,246	17.7	290	14.3
Total	7,031		2,030	

Note. Race data were unavailable for 72 and 50 of the applicants from the written tests and performance tests samples, respectively. Gender or age data were unavailable for seven and 30 applicants, respectively, from the performance tests sample.

All three tests were administered in multiple choice formats with between 70 and 100 items each. The internal consistency reliability estimates (alpha) for the verbal ability, job knowledge, and biodata tests were .90, .92, and .95, respectively.

Performance tests. The five performance tests included three types of interviews (i.e., behavior description, situational, and an experience and interest interview), a leaderless group discussion, and a case analysis exercise. For the behavior description interview, there were five to seven questions that asked candidates to describe examples from their past experiences that demonstrated the specific knowledge, skills, and abilities important for the job. For the situational interview, there were six total questions (three questions for each of two hypothetical job-relevant scenarios), again designed to demonstrate the specific knowledge, skills, and abilities important for the job. The experience and interest interview assessed educational background, work experience, and motivation for and interest in the position using somewhat less structured questions. That is, interviewers chose one question from each of three pairs of questions, for a total of three questions. Evaluation of candidate responses in the experience and interest interview was not based on specific dimensions as in the other two interviews but rather on an overall evaluation.

In the leaderless group discussion, groups of three to six candidates received both common and unique information regarding a set of proposals (with each candidate having to lobby for their particular proposal) and were instructed to reach consensus on the proposal. Finally, in the case analysis exercise, candidates were asked to read and analyze a business case (approximately 10 pages in length) and then write a memo detailing their analysis and recommendations. Each of the performance tests lasted approximately 1 hour.

Candidate performance in each performance test was rated by two assessors, both of whom were experienced in and knowledgeable about the content of the job. In addition, all assessors received 2 days of intensive training on how to rate performance in these assessments. Within each test, assessors rated candidates on multiple dimensions using behaviorally anchored scales. These ratings were then averaged to create an overall performance score for each candidate. We focus on the overall scores for each assessment, for two reasons. First, the ratings in each performance test can be best described by a single overall performance dimension, which is common in both interviews and assessment centers (e.g., Lance, 2008; Sackett & Dreher, 1982; Van Iddekinge, Raymark, Eidson, & Attenweiler, 2004). Second, the organization uses the mean test scores, rather than the dimension scores, to make selection decisions. The reliability estimates for performance test scores (averaged across the two raters) appear in Table 2.

Test development. The tests were designed to be equivalent across testing administrations, following best practices in test construction and professional standards (e.g., Society for Industrial and Organizational Psychology, 2003). For the written tests, there is a detailed blueprint that dictates the test content (e.g., the number of items per topic area). The process for item writing for each administration is the same; each assessment undergoes the same content validation process using subject matter experts from the same population of job experts, and all new forms are extensively pilot tested. In addition, item statistics are used for decisions regarding item inclusion on each exam, thus helping to ensure comparability in difficulty across different versions of the same test. For the performance tests, there are multiple parallel versions of each assessment. They are parallel with regard to number of items, dimensions tapped, length, and reading level. As with the written tests, the performance tests are also extensively reviewed by subject matter experts and pilot tested.

To assess this equivalence, we compared scores for the Time 1 written and performance tests across each administration period. The means and standard deviations were highly similar across administrations, which helps rule out the possibility that any retest effects we may observe are merely a function of differences in test difficulty.

Feedback to applicants. With respect to feedback that applicants received, they were told their composite score on both the written and performance tests, as well as the cut score for each set of tests. Thus, applicants knew how close they were to the overall passing score. They were not told their scores on each specific assessment, but they could ask in writing to see these separate scores (typically only about 10% of these applicants make such a request). To assess whether those applicants who decided to retest were primarily clustered around the cut score, we examined frequency distributions of composite scores for both the written and performance tests. In both cases, although the distributions indicated that a large number of retesters were within a few points of the score cutoff, the majority (around two thirds in both cases) were not.

Results

General Findings

We begin by discussing some general findings that emerged from the analyses. Table 2 presents correlations between initial and retest scores for all eight selection tests. There are two noteworthy

Table 2
Observed Correlations Between Initial and Retest Scores

Test occasion/assessment	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Initial scores for written tests																
1. Verbal ability test	.90															
2. Job knowledge test	.57	.92														
3. Biodata test	−.07	−.14	.95													
Retest scores for written tests																
4. Verbal ability test	.81	.58	−.04	.90												
5. Job knowledge test	.59	.71	−.09	.62	.92											
6. Biodata test	.04	−.02	.69	.07	.02	.95										
Initial scores for performance tests																
7. Behavior description interview	.04	.06	.11	.03	.02	.09	.84									
8. Situational interview	.04	.08	.03	.03	.06	.01	.62	.86								
9. Experience and interest interview	.02	.00	.07	−.01	.00	.06	.57	.49	.83							
10. Leaderless group exercise	.11	.06	−.03	.10	.07	−.03	.15	.14	.13	.87						
11. Case analysis exercise	.21	.16	−.10	.20	.12	−.10	−.04	−.02	−.08	.06	.85					
Retest scores for performance tests																
12. Behavior description interview	.03	−.02	.07	.07	−.04	.10	.30	.20	.19	.12	.04	.90				
13. Situational interview	.06	.03	−.01	.09	.04	.03	.23	.25	.16	.18	.05	.65	.89			
14. Experience and interest interview	.05	−.04	.04	.09	−.02	.07	.24	.17	.26	.13	.04	.65	.59	.87		
15. Leaderless group exercise	.14	.06	−.05	.18	.05	.01	.17	.13	.09	.23	.12	.32	.32	.32	.91	
16. Case analysis exercise	.23	.13	−.07	.19	.11	−.03	.15	.14	.06	.14	.23	.23	.27	.24	.37	.92

Note. $N = 7{,}031$ for correlations among the written tests, and 2,060 for correlations among the performance tests and between the written tests and the performance tests. Correlations of an absolute value of .04 and larger are statistically significant, $p < .05$ (two-tailed). Reliability estimates are listed on the diagonal. Estimates for the written tests are coefficient alphas, whereas estimates for the written tests are intraclass correlation coefficients (C, *k*; McGraw & Wong, 1996).

findings concerning these correlations. First, most of the correlations between scores on the different tests were small to moderate in magnitude (Cohen, 1988), ranging from –.14 between the job knowledge test and biodata test (initial test), to .65 between the behavior description interview and both the situational interview and the experience and interests interview (retest). These generally modest relationships provide justification for examining retesting effects separately by test type. Second, initial test–retest correlations for a given test ranged from .69 to .81 for the written tests and from .23 to .30 for the performance tests. These correlations, which do not approach unity, suggest the existence of individual differences in score improvement over time (i.e., changes in relative order), particularly for the performance tests. We return to the practical implications of such differences in the Discussion.

Table 3 presents descriptive statistics and observed and corrected retest effects (i.e., *d*, the standardized mean difference between initial and retest scores) across all applicants for the written and performance test samples. We used a within-persons design (i.e., repeat applicants only) to test our hypotheses and research questions because between-subject designs that compare one-time applicants to retesting applicants can be difficult to interpret given changes in applicant pool characteristics (e.g., quality, size) and selection rates across test occasions (Lievens et al., 2005). The potential for differential drop-out rates across applicant subgroups (Schmit & Ryan, 1997) also can complicate the interpretation of one-time versus retest comparisons.

Observed retest effects ranged from 0.07 to 0.42 across the eight tests, with an overall mean of 0.12. The verbal ability test had the smallest observed retest effect ($d = 0.07$). Larger retest effects were found for the situational interview and case analysis exercise (both ds $= 0.12$), the experience and interest interview ($d = 0.14$), and the job knowledge test and behavior description interview (both ds $= 0.15$). The largest observed retest effects were found for the biodata test ($d = 0.30$) and the leaderless group discussion ($d = 0.42$). These results are consistent with prior research showing that, in general, applicants improve their test scores upon retesting. At the same time, they again suggest the importance of examining retesting effects separately by test type, as the magnitude of score improvement varied by test. Later in the article we report results of some supplementary analyses that may shed light on this pattern of retest effects across test type.

Table 3 also lists retest effects corrected for range restriction and measurement error (unreliability). The specific statistical artifacts that can influence retest effects and the manner in which they should best be addressed have received little attention in the retesting literature. This is particularly true for the issue of range restriction (see Lievens et al., 2005). Although range restriction would be a potential issue in any retesting situation, the multiple-hurdle design of our study required that we give considerable thought to this issue.[3] In the Appendix, we provide a detailed description of and rationale for our approach to correcting for range restriction, including why we report effects corrected for range restriction for the performance tests but not for the written tests.

As can be seen in Table 3, correcting for range restriction increased the retest effects for some of the performance tests (e.g., the behavior description interview) but decreased the effects for others (e.g., the case analysis exercise). One reason for this is that range restriction can affect both the numerator of

[3] We thank Huy Le and Paul Sackett for their invaluable assistance in thinking through the logic and computation of the range restriction corrections.

Table 3
Descriptive Statistics and Retest Effects Across All Applicants in Written and Performance Tests Samples

		Time 1 scores		Time 2 scores		Retest effects		
Sample/assessment	*N*	*M*	*SD*	*M*	*SD*	*d*	d_{c1}	d_{c2}
Written test sample	7,031							
Verbal ability test		47.73	10.20	48.49	10.48	0.07		0.08
Job knowledge test		46.80	9.27	48.17	9.49	0.15		0.15
Biodata test		45.63	8.74	48.46	9.91	0.30		0.31
Performance test sample	2,060							
Behavior description interview		5.07	0.54	5.16	0.59	0.15	0.27	0.30
Situational interview		4.85	0.60	4.92	0.63	0.12	0.20	0.22
Experience and interest interview		5.15	0.57	5.23	0.65	0.14	0.16	0.18
Leaderless group exercise		4.64	0.53	4.88	0.58	0.42	0.40	0.44
Case analysis exercise		4.53	0.67	4.61	0.73	0.12	0.05	0.05

Note. d = observed standardized mean difference [$(M_{\text{Time }2} - M_{\text{Time }1})/SD_{\text{pooled}}$]; d_{c1} = d corrected for indirect range restriction due to preselection on the written tests; d_{c2} = d corrected for range restriction and measurement error (estimated using alpha coefficients for the written tests and interrater coefficients for the performance tests) in Time 1 scores. Written test scores were not corrected for range restriction (see the Appendix for a detailed explanation of the range restriction corrections).

d values (i.e., the mean differences) and the denominator (i.e., the standard deviations). As a result, the cumulative effect of the corrections can be difficult to predict. For example, the corrections sometimes lowered the Time 1 mean on the performance tests (which, all things being equal, increases *d*) but simultaneously increased the variance in scores (which, all things being equal, decreases *d*). Thus, unlike correcting validity correlations for range restriction—which almost always increases the correlations—correcting retest effects for range restriction may not always increase the observed effects.

Tests of Hypotheses

Table 4 presents observed retest effects by applicants' demographic subgroup. Our hypotheses concerned whether candidates from certain subgroups improved more upon retesting than did other candidates. We tested these hypotheses via multiple regression analyses, whereby Time 2 test scores were regressed on Time 1 scores and a dummy-coded variable that reflected candidates' standing on the demographic variable of interest. A statistically significant beta weight for the demographic variable provides evidence that there are demographic differences in retesting score gains. Results from these analyses are given in Table 5.

Hypothesis 1 predicted that there would be larger score gains with retesting for White applicants than for minority (Black and Hispanic) applicants. As Table 5 indicates, this hypothesis was supported for all three of the written tests. As the associated retest effects in Table 4 show, Whites demonstrated significantly larger retesting score gains on the verbal ability test ($d = 0.12$) compared with Blacks ($d = 0.04$), Hispanics ($d = 0.06$), and Asians ($d = 0.04$). Similar patterns were obtained for both the job knowledge test and the biodata test (see Tables 4 and 5).

With regard to the performance tests, Table 5 indicates that the only performance test on which Whites showed significantly larger retesting score gains was the case analysis, where they showed larger gains than Hispanics (*d*s = 0.11 vs. 0.07, respectively). In fact, on the three interviews the pattern was reversed in several comparisons, with Blacks showing significantly larger retesting gains than Whites on each of the three interviews. Further, Asians

Table 4
Observed Retest Effects by Applicant Subgroup and Overall in Written and Performance Test Samples

	Written test sample			Performance test sample				
Subgroup	VA	JK	BIO	BDI	SI	EI	LG	CA
White	0.12*	0.21*	0.37*	0.14*	0.12*	0.14*	0.43*	0.11*
Black	0.04	0.08*	0.17*	0.25*	0.34*	0.32*	0.37*	0.26*
Hispanic	0.06*	0.10*	0.22*	0.09	0.02	0.00	0.40*	0.07
Asian	0.04	0.08*	0.24*	0.21*	0.21*	0.27*	0.43*	0.15
Male	0.08*	0.13*	0.31*	0.10*	0.09*	0.10*	0.40*	0.02
Female	0.07*	0.18*	0.30*	0.23*	0.18*	0.24*	0.47*	0.31*
Under 40	0.10*	0.18*	0.33*	0.18*	0.16*	0.19*	0.44*	0.12*
40 and older	0.01*	0.03	0.19*	−0.06	−0.09	−0.08	0.32*	0.09
All applicants	0.07*	0.15*	0.30*	0.15*	0.12*	0.14*	0.42*	0.12*

Note. VA = verbal ability test; JK = job knowledge test; BIO = biodata test; BDI = behavior description interview; SI = situational interview; EI = experience and interests interview; LG = leaderless group exercise; CA = case analysis exercise. Statistics reflect *d* values uncorrected for range restriction or measurement error.
[*] Mean difference associated with *d* (Time 2 − Time 1) is statistically significant, $p < .05$ (two-tailed).

Table 5
Results of Multiple Regression Analyses Testing the Effects of Applicant Demographics on Retest Score Improvement

	White–Black			White–Hispanic			White–Asian			Gender			Age		
Assessment/predictor	*B*	*SE*	β	*B*	*SE*	β	*B*	*SE*	β	*B*	*SE*	β	*B*	*SE*	β
Verbal ability test															
Time 1 scores	0.80	0.01	.78	0.80	0.01	.78	0.80	0.01	.78	0.83	0.01	.81	0.80	0.01	.78
Demo variable	2.41	0.24	.08*	1.45	0.24	.04*	1.58	0.23	.05*	−0.14	0.15	−.01	2.28	0.20	.08*
Job knowledge test															
Time 1 scores	0.69	0.01	.67	0.69	0.01	.67	0.69	0.01	.67	0.72	0.01	.71	0.72	0.01	.70
Demo variable	2.93	0.25	.10*	2.22	0.26	.07*	2.23	0.24	.08*	0.41	0.16	.02	2.02	0.21	.08*
Biodata test															
Time 1 scores	0.78	0.01	.69	0.78	0.01	.69	0.78	0.01	.69	0.78	0.01	.69	0.78	0.01	.69
Demo variable	1.48	0.26	.05*	0.98	0.28	.03*	0.71	0.26	.02*	−0.06	0.17	.00	1.13	0.22	.04*
BD interview															
Time 1 scores	0.32	0.02	.30	0.32	0.02	.30	0.32	0.02	.30	0.31	0.02	.29	0.32	0.02	.29
Demo variable	−0.14	0.07	−.04*	−0.02	0.05	−.01	−0.03	0.05	−.01	−0.14	0.03	−.11*	0.08	0.04	.05*
S interview															
Time 1 scores	0.26	0.02	.25	0.26	0.02	.25	0.26	0.02	.25	0.26	0.02	.25	0.26	0.02	.25
Demo variable	−0.17	0.08	−.05*	0.06	0.06	.02	−0.01	0.05	−.01	−0.11	0.03	−.08*	0.09	0.04	.05*
EI interview															
Time 1 scores	0.29	0.02	.25	0.29	0.02	.25	0.29	0.02	.25	0.28	0.02	.25	0.29	0.02	.26
Demo variable	−0.18	0.08	−.05*	−0.01	0.06	−.00	−0.14	0.05	−.06*	−0.15	0.03	−.11*	0.13	0.04	.07*
Leaderless group															
Time 1 scores	0.25	0.02	.23	0.25	0.02	.23	0.25	0.02	.23	0.25	0.02	.23	0.25	0.02	.23
Demo variable	−0.02	0.07	−.01	0.06	0.05	.02	0.01	0.05	.01	−0.07	0.03	−.06*	0.09	0.04	.05*
Case analysis															
Time 1 scores	0.22	0.02	.21	0.22	0.02	.21	0.22	0.02	.21	0.25	0.02	.23	0.23	0.02	.21
Demo variable	0.05	0.09	.01	0.27	0.07	.09*	0.00	0.06	.00	−0.19	0.03	−.12*	0.10	0.04	.05*

Note. BD = behavior description; S = situational; EI = experience and interests; Demo variable = relevant demographic variable. Regression coefficients for all Time 1 scores are statistically significant ($p < .05$). Demographics were coded such that positive regression coefficients reflect larger score improvement for majority group applicants (i.e., Whites, Males, and those under 40 years of age). All three racial group comparisons were included in a single regression model, and thus the statistics for Time 1 scores are the same across the relevant columns.
* $p < .05$ (two-tailed).

showed significantly larger retesting gains than Whites on the experience and interests interview.

Overall, support for Hypothesis 1 was consistent across the three written tests, indicating that Whites demonstrated significantly larger score gains on each of those tests compared with both Blacks and Hispanics (and Asians). In addition, the fact that this pattern did not hold for the performance tests provides support for Hypothesis 2, which predicted that the differential retesting gains in favor of White applicants would be greater on the written tests than on the performance tests. In fact, in some instances (namely, the interviews), minority applicants (particularly Blacks) showed significantly larger increases with retesting compared with White applicants.

Hypothesis 3 predicted that there would be larger score gains for applicants under 40 than for applicants 40 and over. As Table 5 indicates, this hypothesis was supported across all eight tests. Applicants under 40 showed significantly larger score gains than those over 40 on the verbal ability test, the job knowledge test, the biodata test, the three interviews, the leaderless group discussion, and the case analysis exercise (all *p*s $< .05$). Differences in retest effects between the younger and older applicants ranged from *d*s of 0.10 versus 0.01, respectively, for the verbal ability test to 0.19 versus −0.08, respectively, for the experience and interest interview (see Table 4).

We also explored possible differences in retest effects with respect to gender. As Table 5 indicates, there were no significant gender effects on any of the written tests. Thus, for those tests, it appears that men and women gain approximately the same amount with retesting. However, an interesting and consistent pattern emerged across the performance tests. For all five of these tests, women showed significantly larger score gains with retesting compared with men. The differences in retest effects between women and men ranged from *d*s of 0.47 versus 0.40, respectively, for the leaderless group discussion, to 0.31 versus 0.02, respectively, for the case analysis exercise (see Table 4).

Supplemental Analyses

The pattern of differences in retest effects by test type raises the important question of what it is about these various tests that can explain (a) why some tests (e.g., biodata and the leaderless group discussion) reveal larger retesting gains than others (e.g., verbal ability) and (b) why demographic differences in these gains appear to vary across test type (e.g., Whites have a retesting advantage primarily on written tests; women have a retesting advantage only on performance tests). We undertook a post hoc examination of several possibilities, based on the notion that selection tests (including the ones used here) vary in a number of ways that might help explain our pattern of observed results. In interpreting these results, however, it is important to emphasize that they are preliminary in nature and primarily intended to stimulate future research.

We identified five features on which these tests differ that might be relevant to retest effects in general and/or to demographic differences in retest effects in particular. These features include the extent to which the test is (a) g-loaded (i.e., cognitive ability loaded), with the rationale being that g-loaded tests should show minimal score changes over time because ability is a relatively stable individual difference (Lievens et al., 2005); (b) novel, with the rationale being that larger score gains may be obtained with more novel tests, because there is more to be learned from the first exposure to a more novel test than a less novel test; (c) fakeable, with the rationale being that tests that are more fakeable might show larger improvements with retesting, because score gains on such tests could be due to deliberate response distortion rather than (or in addition to) true improvement on the construct of interest; (d) based on identical versus parallel forms across administrations, with the rationale being that identical tests generally show larger improvement with retesting than parallel tests (Hausknecht et al., 2007); and (e) sign- versus sample-oriented (Schmitt & Ostroff, 1986; Wernimont & Campbell, 1968), given the argument presented earlier that the differential retesting gains in favor of White applicants may be smaller for samples (i.e., tests closer to job performance) than for signs.

To empirically assess whether these five test features could explain the pattern of retest effects obtained across demographic groups and tests, we first independently rated each of the eight tests on each of these five dimensions using 3-point scales (low, medium, high) for all of the dimensions except identical versus parallel, which was dichotomous. The mean intraclass correlation coefficient (C, k) across these ratings was .89. We then averaged the ratings across the four raters to obtain a mean score on each dimension for each of the eight tests. These scores are reported in the top half of Table 6.

To assess whether these dimensions could explain variance in retest effects across the tests, we computed correlations between these dimension ratings and the retest effects (d) across the eight tests. These correlations are presented in the bottom half of Table 6 for both the overall ds and the ds for each demographic subgroup. Two points should be kept in mind when interpreting these results. First, these correlations are based on only eight data points (one for each test). Second, the performance tests sample is a subset of the written tests sample. Thus, the retest effects for the three written tests and five performance tests are based on overlapping but different samples.

As Table 6 shows, ratings of novelty (r = .65), g-loading (r = –.46), sign–sample (r = .42), and fakeability (r = .36) were correlated with the overall retest effects. This suggests that larger retest score gains overall were observed with more novel tests, less g-loaded tests, sample-based tests, and more fakeable tests. However, as expected, these relationships also differed somewhat across demographic subgroups, suggesting that these test dimensions may better explain the pattern of retest effects for some groups than for others. As one example, consider the novelty dimension. Although the overall correlation suggests that the novelty of a test may be positively related to test score gains for all groups (with larger gains seen for more novel tests), the subgroup correlations are notably larger for women than for men (rs = .83 vs. .48) and for those over 40 compared with those under 40 (rs =

Table 6
Results of Supplemental Analyses Relating Assessment Dimensions to Retest Effects

Assessment	Assessment dimension mean ratings: Novelty	Sign–sample	g loading	Fakeability	Parallel vs. identical forms
Verbal ability test	1.00	1.25	3.00	1.00	1
Job knowledge test	1.25	1.50	3.00	1.00	1
Biodata test	2.50	1.25	1.00	3.00	2
Behavior description interview	1.75	2.00	1.75	2.00	1
Situational interview	1.75	2.00	2.00	2.00	1
Experience and interests interview	1.00	1.50	1.00	2.75	2
Leaderless group exercise	3.00	3.00	1.75	1.75	1
Case analysis exercise	3.00	2.75	2.25	1.25	1

Candidate group	Correlations between assessment dimension ratings and retest effects effects (d): Novelty	Sign–sample	g loading	Fakeability	Parallel vs. identical forms
All candidates	.65	.42	−.46	.36	.19
White candidates	.54	.21	−.36	.32	.25
Black candidates	.45	.67	−.61	.45	.08
Hispanic candidates	.66	.46	−.19	.08	−.05
Asian candidates	.54	.55	−.71	.56	.26
Male candidates	.48	.20	−.40	.37	.24
Female candidates	.83	.73	−.51	.30	.12
Under 40 candidates	.56	.34	−.50	.43	.26
40 and over candidates	.75	.45	−.13	.01	.02

Note. The first four dimensions were rated on a 3-point scale, with higher numbers indicating tests were more novel, more sample-like, more g-loaded, and more fakeable. For the fifth dimension, 1 = parallel forms and 2 = identical forms used across administrations. The rs at the bottom of the table are zero-order correlations between assessment dimension ratings and retest effects (d) across the eight tests (thus N = 8), computed for both all candidates and each demographic subgroup (corresponding d values are listed in Table 4).

.75 vs. .56). In other words, the novelty of the test appears more important in explaining retest effects across tests for these minority applicants than for their majority counterparts. Another example is provided by the sign–sample dimension. Although the overall correlation of .42 suggests that larger score gains are observed for more sample-based (as opposed to sign-based) tests, the correlations are notably larger for racial minorities (Black, Hispanic, and Asian applicants) than for White applicants, and for female applicants compared with male applicants.

Discussion

In this article we sought to examine retest effects across a wide range of assessments to answer the important question of whether different demographic groups show differential levels of score improvement. Supporting the relevance of exploring such differences, Hausknecht et al.'s (2007) meta-analysis revealed evidence of significant heterogeneity in retest effects across studies, suggesting the existence of moderators (earlier reviews by Kulik et al., 1984, and Sackett et al., 1989 reached similar conclusions). In addition, Reeve and Lam (2007) found significant variance in score gains across people and tests even within the same setting, leading them to encourage more research on relations between examinee characteristics and retest effects. Although there may be multiple characteristics responsible for such heterogeneity, we focused on demographic variables, given the important practical and legal implications noted previously (i.e., adverse impact) as well as emerging theoretical arguments about the role of test attitudes and ability in retest effects. With this focus as a backdrop, we summarize the four primary conclusions of our study and their corresponding contributions to the literature.

Race Differences in Retesting Effects

First, we examined whether retesting effects differed across racial groups, given the obvious implications for adverse impact. That is, if minority applicants improve more with retesting, it suggests that retesting programs may have the added benefit of reducing adverse impact. On the other hand, if White applicants improve more with retesting, such policies might exacerbate levels of adverse impact. Ployhart and Holtz (2008) recently discussed retesting policies in their review of strategies designed to reduce racial and gender-based adverse impact. Although they concluded that such policies were not effective in reducing adverse impact, the only available empirical study on which this conclusion was based is a conference paper (Sin, Farr, Murphy, & Hausknecht, 2004) that used a between-subjects design (see Lievens et al., 2005). In addition, the possibility that Whites could show larger increases with retesting has not been addressed.

Our results show that Whites did, in fact, evidence significantly larger score improvements with retesting than Blacks and Hispanics on a number of selection tests, including tests of job knowledge, biodata, and verbal ability. Whites also showed significantly larger score improvements than Hispanics on the case analysis exercise. For the interviews, however, the pattern was reversed, with Blacks showing significantly larger retesting gains than Whites on each of the three interviews.

This pattern of findings supports our race-based hypotheses and suggests that differences in both ability and testing attitudes and beliefs are viable theoretical explanations for racial differences in retest effects. The fact that the differential gains of Whites over minorities were more pronounced with the written tests further supports the viability of both theoretical explanations, given that it is the written tests for which (a) negative testing attitudes on the part of minorities would likely be most salient and (b) g-saturation is greatest, leading to the differential ability advantage for White applicants. It is important that both theoretical processes are likely to be manifest in many employee selection contexts, suggesting there is good reason to believe these demographic differences in retesting gains may generalize across organizations and jobs. Although we could not directly examine the role of either of these theoretical explanations in this study, we point to them as two viable areas for future research to examine.

Age Differences in Retesting Effects

Second, we predicted that retesting effects would also differ across age groups (under 40 vs. 40 and over). We found consistent support for this hypothesis, in that applicants under the age of 40 demonstrated significantly larger score gains, compared with those 40 and older, on all eight tests. This finding may be due to the possibility that older test takers have less positive attitudes toward testing in general and their own self-efficacy for doing well on tests (e.g., Ackerman et al., 2002; Hausknecht et al., 2004), and/or age-related declines in test-taking skills (Sarnacki, 1979) and fluid intelligence (Kanfer & Ackerman, 2004; Schaie, 1996). It is important to note that the theoretical role of fluid intelligence applies to both performance tests and written tests. Fluid intelligence is the ability to think and reason abstractly and to solve problems; it is generally considered independent of learning, experience, and education (Cattell, 1987; Horn & Cattell, 1967). This is likely to be important for improving on all selection tests, particularly those that have not been previously encountered by applicants. Again, future research should attempt to isolate the importance of these two theoretical processes (i.e., age-related changes in testing motivation and fluid intelligence) in age differences in retesting score gains. Given the robust pattern of findings and the associated implications for age-related adverse impact (which we demonstrate under Practical Implications below), this is certainly an area worthy of further study.

Gender Differences in Retesting Effects

Third, we found that the size of retesting score gains also differed on the basis of gender, such that women showed significantly larger score gains with retesting than men on all five performance tests. These findings were somewhat unexpected in that there was no strong empirical evidence base or theoretical arguments for expecting gender differences in either ability or testing attitudes or motivation.

One possible contributing factor might be related to gender differences in receptivity to feedback. Being told that one has failed a selection test (as the retesters in the current study were) constitutes negative feedback, and research has shown that women tend to react more positively than men when they receive negative feedback and are also more likely to make use of that feedback (Johnson & Helgeson, 2002). For example, women tend to perceive such feedback as being more accurate and as providing more useful information about themselves (Roberts & Nolen-Hoeksema, 1989, 1994). Research further suggests that women may be more likely to attribute failure (e.g., on a selection test) to internal rather than external factors (e.g., Boggiano & Barrett, 1991; Hirshy & Morris, 2002), which in turn may influence how they approach subsequent test attempts. The fact that in the current study we saw differential retesting gains by women only on the performance tests suggests that it may be these more "novel" tests in which an applicant's feedback orientation is important for determining subsequent test improvement. This is admittedly speculative, but the possible role of gender differences in feedback orientation should be examined in future retesting research.

Retesting Differences Across Type of Assessment

Fourth, we found that retesting effects differed by type of assessment, both overall retest effects (see Table 3) and differential improvement across race and gender subgroups (see Table 4). Regarding the overall retest effects, the largest effects were obtained for the biodata test and the leaderless group discussion, and the smallest effects were obtained for the verbal ability test. Our supplemental analyses suggest that this may be due to the possibility that more novel and more fakeable tests show larger retest gains, whereas more *g*-loaded tests show smaller retest gains. Regarding how race and gender differences varied across type of assessment, the pattern of findings showed larger score gains for White applicants on the verbal ability, job knowledge, and biodata tests, but not on the performance tests. Female applicants' differential improvement, on the other hand, was limited to the performance tests.

Thus, at a general level, our results suggest that retesting effects are not homogeneous across selection tests. Conceptually, there are likely to be different mechanisms underlying retesting improvement across these tests (e.g., actual improvement vs. learning of testing tricks or faking). More research is clearly needed to understand how (and why) retest effects vary across assessments, but in the short term, the current results suggest that researchers should avoid making global statements about retesting and instead confine their conclusions to retest effects on specific assessments. Similarly, our results provide little evidence for the existence of any general "retesting skill," as correlations between applicants' score improvements across different assessments were all less than .20. In short, it does not appear that some people are simply better retesters across the board than others.

Finally, our supplemental analyses suggested that a number of dimensions along which selection tests vary (i.e., novelty, sign–sample, *g* loading, and fakeability) may influence retest effects in general as well as demographic differences in these effects. For example, not only were novelty and the extent to which tests were sample- as opposed to sign-based positively related to overall score gains but also the influence of these factors appeared to be stronger for some minority groups than for the majority groups. With regard to novelty, the positive association between novelty and retest improvement suggests the existence of criterion-relevant changes, whereby a deficit between the observed score and true score at Time 1 (e.g., due to test unfamiliarity or anxiety or stress) is reduced at retesting (Hausknecht et al., 2005; Lievens et al., 2005). It also is interesting to reflect on how this dimension might interact with demographics. Our very preliminary results suggest that with more novel tests, initial test scores may be more deficient indicators of true levels of performance for some minority groups. Thus, allowing retesting might be particularly appropriate with more novel tests (e.g., the leaderless group discussion and case analysis exercise in our study). These possibilities suggest some interesting directions for future retesting research.

Practical Implications

The results of this study have a variety of implications for practice. First, the overall retest effects we discovered could have a notable impact on individual selection decisions. To illustrate, we used hypothetical yet realistic selection ratios of .10, .25, and .50 (e.g., Collins & Morris, 2008) to determine the extent to which the same set of applicants would (or would not) be selected using their initial scores versus their retest scores. For a unit-weighted composite of the written tests, between 15% and 23% of applicants would experience different selection decisions across the two testing occasions under these selection scenarios. The differences were even more severe for selection on a composite of the performance tests, such that between 16% and 39% of applicants would be selected on one occasion, but not on the other.

The differential test score improvements we observed across applicant subgroups also may have notable practical implications, particularly with regard to adverse impact. To illustrate this, we used the same three hypothetical selection ratios as above to determine what adverse impact ratios would be for initial versus retest scores on both the written and performance tests with respect to each of four subgroup comparisons: White–Black, White–Hispanic, male–female, and under 40–40 and older. Results of these analyses are presented in Table 7. Negative values in the table indicate that the adverse impact ratio was smaller upon retesting, which means that the disparity between majority and minority group selection ratios was greater upon retesting (and therefore more problematic from an adverse impact perspective).

For the written tests, there were nine sets of analyses (i.e., three tests by three selection ratios) for each subgroup comparison (see top half of Table 7). With the exception of gender differences on the job knowledge test, adverse impact ratios favoring majority group applicants tended to worsen upon retesting for all subgroup comparisons at all three selection ratios. This effect was the strongest and most consistent for comparisons involving White and Black applicants, such that on average, adverse impact ratios worsened by 0.31 upon retesting.

For the performance tests, there were 15 sets of analyses (i.e., five tests by three selection ratios) for each subgroup comparison (see bottom half of Table 7). Analysis of changes in adverse impact ratios on these tests revealed a more mixed pattern of results. The pattern of results for comparisons between White and Hispanic applicants and between applicants under 40 and 40 and older was similar to the pattern of results for the written tests noted above, such that adverse impact ratios (favoring Whites and younger applicants) worsened with retesting. However, the opposite occurred for comparisons between White and Black applicants and between male and female applicants. Specifically, initial adverse impact ratios favoring the minority groups (i.e., Blacks and women) tended to improve upon retesting.

The above results illustrate that even seemingly small differential retest effects (which in our case tended to favor majority

Table 7
Hypothetical Changes in Adverse Impact Ratios Based on Initial Test Scores and Retest Scores

Assessment/selection ratio	White–Black	White–Hispanic	Male–female	Under 40–40 and older
Verbal ability test				
.10	−0.09	0.05	−0.00	−0.01
.25	−0.01	−0.00	−0.06	−0.01
.50	−0.01	−0.01	−0.04	−0.02
Job knowledge test				
.10	−0.37	0.01	0.08	−0.24
.25	−0.45	−0.01	0.08	−0.17
.50	−0.48	−0.04	0.06	−0.07
Biodata test				
.10	−0.64	−0.50	−0.16	−0.32
.25	−0.31	−0.17	−0.11	−0.18
.50	−0.10	−0.07	−0.05	−0.03
Written tests overall	−0.31	−0.01	−0.04	−0.07
BD interview				
.10	−0.18	0.52	0.30	−0.50
.25	−0.28	0.07	0.12	−0.25
.50	0.08	−0.20	0.12	−0.20
Situational interview				
.10	0.70	−0.38	0.40	−0.37
.25	0.02	−0.33	0.16	−0.29
.50	0.12	−0.11	−0.03	−0.08
EI interview				
.10	0.48	−0.02	0.15	−0.30
.25	0.47	−0.12	0.14	−0.22
.50	0.16	−0.10	0.01	−0.12
Leaderless group exercise				
.10	−0.05	−0.46	−0.09	−0.24
.25	0.15	−0.13	0.28	0.11
.50	0.15	−0.10	0.09	0.01
Case analysis exercise				
.10	0.36	0.25	0.17	0.07
.25	0.37	−0.25	0.43	0.05
.50	0.29	−0.06	0.28	0.05
Performance tests overall	0.15	−0.11	0.15	−0.20

Note. Values reflect the change in adverse impact (AI) ratios for the two relevant groups based on initial test scores versus retest scores (i.e., $AI_{Retest} - AI_{Initial}$). Negative values indicate that the AI ratio was smaller upon retesting, which means that AI (or the disparity between majority and minority group selection ratios) was worse upon retesting. Values for written tests overall and performance tests overall reflect the median across tests and selection ratios. BD = behavior description; EI = experience and interests.

groups for the written tests and some minority groups for the performance tests) can influence adverse impact, which, in turn, can have consequences for diversity and legal defensibility. In addition, changes in adverse impact ratios from initial test to retest tended to be largest at lower selection ratios (the mean absolute changes in adverse impact across tests and groups were .26, .18, and .19 for the selection ratios of .10, .25, and .50, respectively), which suggests that retesting policies may have the greatest influence on adverse impact in highly competitive selection contexts. In addition, we note that differential drop-out rates across racial groups (Schmit & Ryan, 1997), which we do not consider here, could further exacerbate changes in adverse impact with retesting.

Limitations and Directions for Future Research

We conclude by noting some potential limitations of the current study, as well as some possible directions for the future. First, our sample primarily comprised highly educated candidates applying for relatively high-level "knowledge work" positions. Although one could view this as a strength from a contribution perspective (extending previous retesting findings to a new context), it could be simultaneously viewed as a limitation from a generalizability perspective. Thus, future research should continue to examine retesting effects on actual job applicants, across a variety of jobs. Second, special attention should be paid to including diverse samples in future retesting research. Although our samples were relatively diverse, the number of Black applicants retaking the performance tests was relatively small ($n = 65$).

Third, although we examined a larger range of selection tests than previous retesting research has examined, it was not an exhaustive list. For example, with the exception of the biodata test, none of the assessments we examined were likely to be highly susceptible to applicant response distortion. Thus, future research might incorporate assessments such as personality inventories and integrity tests along with more objective tests to investigate the extent to which impression management processes might help explain the pattern of retest effects across assessments.

Fourth, although we offered preliminary theoretical explanations for the demographic differences in retest effects (including differences in ability, testing attitudes and motivation, and receptivity to feedback), we did not measure these constructs and thus could not directly assess their potential influence. A critical need for future research is to incorporate direct measures of these potential processes in order to isolate the *why* of retest effects, including why there may be demographic differences in such effects. In the meantime, we see value in having specifically identified several theoretically relevant variables here, to help guide this future research. It is important to note that, as our discussion shows, there is no reason to believe that the same theoretical explanation would apply across demographic groups. That is, differences in ability, testing attitudes and motivation, and feedback orientation may be differentially relevant for explaining race versus age versus gender differences in retesting effects. In short, "one size does not [necessarily] fit all" when it comes to explaining demographic differences in retesting score improvements.

Fifth, we did not have access to performance data on applicants who ultimately were selected. Thus, we were unable to assess the effects of retesting on criterion-related validity. In fact, we are not aware of any published studies examining how retesting affects the prediction of actual job performance in job applicant settings. The results of such research have obvious implications for retesting policies, and the lack of validity research represents a critical gap that future research needs to address. We also did not address the question of *who* retests in the first place. It would be interesting for future research to identify applicant (e.g., demographic and personality variables) and test score (e.g., how close the score is to the cutoff) characteristics that predict the decision to retest.

A final area for future research involves the need to examine other characteristics responsible for heterogeneity in retest effects beyond the demographic characteristics studied here. Some possible candidates for further study include the extent to which applicants are motivated to be selected and relevant personality characteristics such as achievement motivation or conscientiousness. However, given the current findings, we reiterate the need to specify and research for what types of tests these other characteristics would be relevant.

References

Ackerman, P. L., Beier, M. E., & Bowen, K. R. (2002). What we really know about our abilities and our knowledge. *Personality and Individual Differences, 33,* 587–605.

Arvey, R. D., & Sackett, P. R. (1993). Fairness in selection: Current developments and perspectives. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations* (pp. 171–202). San Francisco, CA: Jossey-Bass.

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

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

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.

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.

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.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Collins, M. W., & Morris, S. B. (2008). Testing for adverse impact when sample size is small. *Journal of Applied Psychology, 93,* 463–471.

Duehr, E. E., Jackson, H. L., & Ones, D. S. (2004). *Gender differences in Big Five factors and facets: A meta-analysis.* Unpublished manuscript.

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.

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.

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

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

Gilliland, S. W., & Steiner, D. D. (2001). Causes and consequences of applicant fairness. In R. Cropanzano (Ed.), *Justice in the workplace* (Vol. 2, pp. 175–195). Mahwah, NJ: Erlbaum.

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.

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.

Hausknecht, J. P., Halpert, J. A., Harder, B. K., Kuljanin, G., & Moriarty, M. (2005, April). *Issues in repeated testing: Test attitudes and applicant reactions.* Paper presented at 20th Annual Conference of the Society for Industrial and Organizational Psychology, Los Angeles, CA.

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.

Hirshy, 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.

Hogan, J., Barrett, P., & Hogan, R. (2007). Personality measurement, faking, and employment selection. *Journal of Applied Psychology, 92,* 1270–1285.

Horn, J. L., & Cattell, R. B. (1967). Age differences in fluid and crystallized intelligence. *Acta Psychologica, 26,* 107–129.

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection and amelioration of adverse impact in personnel selection procedures: Issues, evidence and lessons learned. *International Journal of Selection and Assessment, 9,* 152–194.

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

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

Johnson, J. T., & Ree, M. J. (1994). A Pascal program to compute the multivariate correction for range restriction. *Educational and Psychological Measurement, 54,* 693–695.

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

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.

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.

Lance, C. E. (2008). Why assessment centers do not work the way they are supposed to. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1,* 84–97.

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.

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.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1,* 30–46.

McKay, P., & McDaniel, M. A. (2006). A re-examination of Black–White mean differences in work performance: More data, more moderators. *Journal of Applied Psychology, 91,* 531–554.

Nguyen, H.-H. D., & Ryan, A. M. (2008). Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. *Journal of Applied Psychology, 93,* 1314–1334.

Ployhart, R. E., & Holtz, B. C. (2008). The diversity–validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology, 61,* 153–172.

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.

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.

Ree, M. J., Carretta, T. R., Earles, J. A., & Albert, W. (1994). Sign changes when correcting for range restriction: A note on Pearson's and Lawley's selection formulas. *Journal of Applied Psychology, 79,* 298–301.

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.

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.

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

Roberts, T. A., & Nolen-Hoeksema, S. (1994). Gender comparisons in responsiveness to others' evaluations in achievement settings. *Psychology of Women Quarterly, 18,* 221–240.

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.

Roth, P. L., Huffcutt, A. I., & Bobko, P. (2003). Ethnic group differences in measures of job performance: A new meta-analysis. *Journal of Applied Psychology, 88,* 694–706.

Ryan, A. M. (2001). Explaining the Black–White test score gap: The role of test perceptions. *Human Performance, 14,* 45–75.

Saad, S., & Sackett, P. R. (2002). Investigating differential prediction by gender in employment-related personality measures. *Journal of Applied Psychology, 87,* 667–674.

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

Sackett, P. R., & Dreher, G. F. (1982). Constructs and assessment center dimensions: Some troubling findings. *Journal of Applied Psychology, 67,* 401–410.

Sarnacki, R. E. (1979). An examination of test-wiseness in the cognitive test domain. *Review of Educational Research, 49,* 252–279.

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

Schleicher, D. J., Venkataramani, V., Morgeson, F. P., & Campion, M. A. (2006). So you didn't get the job. Now what do you think? Examining opportunity-to-perform fairness perceptions. *Personnel Psychology, 59,* 559–590.

Schmit, M. J., & Ryan, A. M. (1997). Applicant withdrawal: The role of test-taking attitudes and racial differences. *Personnel Psychology, 50,* 855–876.

Schmitt, N., & Ostroff, C. (1986). Operationalizing the "behavioral consistency" approach: Selection test development based on a content-oriented strategy. *Personnel Psychology, 39,* 91–108.

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

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

Uniform Guidelines on Employee Selection Procedures (1978), 43 Fed. Reg. 38290.

Van Iddekinge, C. H., Raymark, P. H., Eidson, C. E., Jr., & Attenweiler, B. (2004). What do structured interviews really measure? The construct validity of behavior description interviews. *Human Performance, 17,* 71–93.

Wernimont, P. F., & Campbell, J. P. (1968). Signs, samples, and criteria. *Journal of Applied Psychology, 52,* 372–376.

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.

Appendix

Corrections for Range Restriction

Range restriction (RR) is a potential issue in any retesting situation, but the multiple-hurdle design of the current study (wherein passing scores on the written tests are required before moving on to take the performance tests) required that we give particularly careful thought to this issue. Unfortunately, the nature of and corrections for range restriction have received scant attention in the retesting literature (see Lievens et al., 2005). For this reason, we provide a detailed description of and rationale for each step in our approach to correcting for range restriction.

We first considered what the relevant population would be for each sample of retesters. For the written tests, applicant retest scores are subject to direct RR due to previous selection on these tests at Time 1. That is, candidates scoring in the top 40% or so passed the written tests and thus are not represented in the retest sample for the written test. Thus, it could be argued that *d* values indexing score differences between initial test and retest are downwardly biased due to the potential restricted variance in retest scores. However, the population involved in a correction for this sort of direct RR would be all applicants who entered the selection process at Time 1. Because our interest was in estimating score changes among applicants who retest, rather than applicants in general (which also would include those who passed and thus did not have to retest), we did not correct the written tests for this form of RR, because doing so would likely overestimate the magnitude of the retest effects.

For the performance tests, there are several potential sources of RR. The first is analogous to the direct RR noted above for the written tests, in the sense that the performance retest (Time 2) scores may have restricted variance because the highest scoring candidates passed the performance tests the first time and therefore are not represented in the retest sample. For the reasons noted above, we did not correct for this "naturally occurring" type of RR, given that our interest is in estimating score changes among applicants who retest, rather than applicants in general.

The second type of RR potentially affecting the performance tests stems from the fact that these assessments were administered as a second hurdle in the selection process. All things being equal, the larger the correlations between written and performance test scores, the more the range of performance test scores will be reduced due to preselection on the written tests. Although our data revealed only modest correlations between the written and performance test scores (see Table 2), our design makes this a crucial form of RR to consider. In short, the performance test scores (both initial and retest) may be subject to indirect RR on the written tests, because only applicants who passed the written tests were invited to take the performance tests. However, as discussed in more detail below, Time 1 and Time 2 performance tests are subject to different forms of this indirect RR, which complicated the required corrections.

Time 1 performance test scores are subject to indirect RR on the Time 1 written tests, because only applicants who passed the written tests were invited to take the performance tests. Time 2 performance test scores are subject to indirect RR on the Time 1 written tests, as well as to indirect RR on the Time 2 written tests. (This is because if a candidate fails the performance tests, they are required to begin the selection process over with the written tests. Thus, anyone repeating the performance tests also would have had to repeat the written tests). Therefore, we wanted to estimate what the retest effects would be if there was no preselection on the written tests (at Time 1 and/or Time 2). Thus, we corrected the Time 1 and Time 2 performance tests scores for RR due to selection on the Time 1 written test scores, and we corrected the Time 2 performance tests for selection on the Time 2 written test scores. The following paragraphs detail the procedure we used for making these corrections.

In the first step, we corrected the Time 2 performance test scores for RR due to preselection on the Time 2 written tests. For this correction, the unrestricted sample comprised applicants who passed the written tests at Time 1, failed the performance tests at Time 1, chose to retest, and retook the written tests (and either passed or failed). The restricted sample comprised applicants who passed the written tests at Time 1, failed the performance tests at Time 1, chose to retest, passed the written tests, and then retook the performance tests (and either passed or failed). Thus, the difference between the unrestricted and restricted samples is that the unrestricted sample included applicants who passed or failed the written tests at Time 2, whereas the restricted sample included only applicants who passed the written tests at Time 2 (and therefore went on to take the performance tests).

To implement this correction, we used the RangeJ software (Johnson & Ree, 1994), which applies the multivariate indirect RR formulas developed by Ree, Carretta, Earles, and Albert (1994). The input for this analysis was the means and standard deviations of the Time 2 written tests in the unrestricted sample and the means, standard deviations, and intercorrelations of the Time 1 and Time 2 written tests and performance tests in the restricted sample.

In the second step, we corrected the Time 1 and Time 2 performance test scores for RR due to preselection on the Time 1 written tests. For this correction, the unrestricted sample comprised all applicants who took the written tests at Time 1 (and either passed or failed), and the restricted sample comprised applicants who passed the written tests at Time 1 and then took and failed the performance tests. Thus, the difference between the unrestricted and restricted samples is that the unrestricted sample included all applicants who took the written tests at Time 1, whereas the restricted sample included only applicants who passed the written tests at Time 1 and then failed the performance tests at Time 1. The input for this analysis was the means and standard deviations of the Time 1 written tests in the unrestricted sample (i.e., all applicants), and the corrected means, standard deviations, and intercorrelations of the Time 1 and Time 2 written tests and performance tests obtained from the first step of the process described above. We then used the resulting corrected means and standard deviations to estimate standardized mean differences (*d*) between Time 2 and Time 1 scores for each performance test.

Note that approximately 39% of applicants passed the written tests at Time 1; of those, 24% took and passed the performance tests at Time 1. Thus, the cumulative selection ratio at Time 1 is approximately 9% (i.e., .39 × .24). However, the means, standard deviations, and correlations we obtained from the above procedure reflect what the values would be if everyone who applied at Time 1 took the performance tests, whereas our population of interest comprises applicants who took and failed the performance tests at Time 1. Unfortunately, we are not aware of a procedure that would enable us to accurately estimate the effects of excluding the 9% of applicants who passed both the written and performance tests at Time 1. Because of this, the RR-corrected retest effects we report (see Table 3) only approximate what the actual effects may be and likely represent conservative estimates of those effects.

Received August 27, 2008
Revision received September 24, 2009
Accepted November 5, 2009 ■