INTELLIGENCE 23, 15–32 (1996)

Short-Term Memory Tests in Personnel Selection: Low Adverse Impact and High Validity

JENNIFER M. VERIVE
MICHAEL A. MCDANIEL
The University of Akron

This study investigated the usefulness of short-term memory tests as a strategy for reducing adverse impact in personnel selection decisions, also achieving high validity in predicting job and training performance. Based on an integration of 27,973 individuals from 31 samples, the average Black–White mean difference in standard deviation units was .42. This is less than half the mean score difference that is typically obtained with general cognitive ability tests. A separate analysis of 141 validity coefficients based on 34,262 individuals indicated that short-term memory tests yielded validities of .41 for job performance (.19 observed) and .49 for training performance (.28 observed). Validity was generalizable for all distributions examined.

Due to the voluntary or court-mandated adoption of affirmative action programs, as well as legal and professional guidelines encouraging the use of selection and promotion tests with the least racial differences, cognitive ability testing has become an area of much interest and concern in industrial and organizational psychology. Whereas most tests of general cognitive ability result in approximately a one standard deviation difference between White and Black mean scores (Gordon, 1986; Herrnstein & Murray, 1994; Jensen, 1985, 1993b; Jensen & Figueroa, 1975; Sattler, 1988; Shuey, 1966; Vincent, 1991; Wonderlic & Wonderlic, 1972), the search for valid tests without this feature is of foremost importance.

Recent attempts to implement selection tools with high validity and minimal adverse impact have followed two approaches. The first and arguably best approach has been to identify additional predictors that would add to the validity of cognitive ability measures and at the same time reduce adverse impact (Hunter & Hunter, 1984). For example, Ones, Viswesvaran, and Schmidt (1993) argued

An earlier version of this article was presented at the 102nd annual convention of the American Psychological Association, Los Angeles, California, August 1994. This paper has benefited from comments provided by Frank L. Schmidt, Arthur R. Jensen, Patrick Kyllonen, Joel P. Wiesen, Douglas K. Detterman, and Peter Legree.

Correspondence and requests for reprints should be sent to Michael A. McDaniel, Department of Psychology, University of Akron, Akron, OH 44325–4301.

that an optimally weighted composite of integrity and cognitive ability tests yields higher validities and less adverse impact than cognitive ability tests alone. The second and less advantageous approach has been to dilute the validity of cognitive ability tests by combining them nonoptimally with less valid predictors or, worse yet, by the abandonment of cognitive ability tests for less valid predictors such as personality inventories (Barrick & Mount, 1991), interviews (McDaniel, Whetzel, Schmidt, & Maurer, 1994), biodata (Rothstein, Schmidt, Erwin, Owens, & Sparks, 1990), or reviews of education and experience (Ash, Johnson, Levine, & McDaniel, 1989; Dye & Reck, 1989; McDaniel, 1986b; McDaniel, Schmidt, & Hunter, 1988a, 1988b).

One type of cognitive ability that has shown promise of filling the need for high validity and low racial differences is short-term memory. Short-term memory testing was popular in the 1960s and 1970s, but has largely been ignored since then (Vernon, 1987). The ability of short-term memory tests to predict job performance with minimal racial differences is explained in Jensen's (1971) level of abilities theory.

Jensen (1971) postulated that cognitive abilities form two levels. Level I abilities are exemplified by short-term memory tasks, such as forward digit span and serial rote learning, which do not require mental manipulation of inputs in order to provide an output. Level II abilities, on the other hand, do require mental manipulation and transformation of inputs in order to arrive at an output or solution. Level II abilities are characterized by the concept of *g*, or general cognitive ability, and are exemplified by performance on *g* tests such as the Raven Progressive Matrices (Raven, Court, & Raven, 1977).

The prototypical Level I test, forward digit span, has been shown to be a separate factor from general cognitive ability (Jensen, 1971). Additionally, Wechsler (1958) noted that forward digit span correlates poorly with other measures of intelligence and that such short-term memory tests are "more or less independent" (pp. 70–71) of general cognitive ability. These results suggest that short-term memory tests may show fewer racial differences than tests that are more highly loaded on general cognitive ability.

Research has provided empirical support for the notion that short-term memory tests result in fewer racial differences than tests of general cognitive ability. A series of studies conducted on elementary school children in northern California found the traditional one standard deviation difference between Black and White participants on intelligence tests (e.g., WAIS), but found less than half a standard deviation difference between Black and White participants on tests of short-term memory (Jensen, 1971, 1974; Jensen & Figueroa, 1975). Further, a literature review summarized by Vernon (1987) concluded that there are fewer differences between low- and high-SES groups on tests of short-term memory.

Although the relatively low level of racial differences in short-term memory tests seems promising, the predictive validity of these tests is less clear. Corey et al. (1980; cited in Jensen, 1993a) gave a battery of tests, including a short-term

memory test, to naval cadets. The authors found that the short-term memory test did predict grades in navy schools (r = .30, corrected for attenuation) and advancement into technical jobs (r = .44, corrected) for naval cadets. However, the short-term memory test was predictive only for cadets with low general cognitive ability. These initial findings suggest that, unlike g, short-term memory tests may not be valid in all situations for all types of individuals.

Recently, Jensen (1993a) revised his position on his original level of abilities theory. Jensen (1993a) stated that his theory is merely an extension of Spearman's hypothesis and thus may not be a unique contribution to the understanding of group differences in general cognitive ability. However, in this same chapter, Jensen (1993a) noted that the distinction between Level I abilities and Level II abilities has unique applied uses in selection and training settings, such as the use of Level I tests to predict training and job success. Further, Jensen (1993a) suggested that there is still need for a "true experimental test on a large enough scale to inspire confidence" (p. 190) in the usefulness of Level I tests. This study attempts to meet this need.

The purpose of this study is twofold. The first goal is to determine if short-term memory tests result in smaller group differences than tests of general cognitive ability. The second goal is to clarify the value of short-term memory tests for predicting job and training performance across jobs and individuals. Both of these goals are addressed using meta-analytic procedures.

METHOD

Two separate meta-analytic studies were conducted to address the two research goals: Study 1 analyzed Black–White differences on short-term memory tests, and Study 2 reviewed the validity of short-term memory tests in predicting job and training performance. Study 1 and Study 2 shared some aspects of methodology, such as decision rules and the literature search. These shared areas are described first, and then the procedures for each study are described separately.

Decision Rules

Rothstein and McDaniel (1989) described the importance of clearly articulated decision rules when conducting and reporting meta-analyses. In order to decide which research studies to include in the meta-analyses and which to omit, two general decision rules were used. In addition to these general rules, Study 1 and Study 2 each had one unique decision rule that guided inclusion of a study.

The first general decision rule, applicable to both Study 1 and Study 2, was that a research study was only included in the meta-analyses if the study used a short-term memory test. Such tests include digit span, digit symbol substitution, or tests very similar to these two tests. Generally, short-term memory tests described in included studies were the digit span and digit symbol subtest of the WAIS, the short-term memory subtests of the Stanford–Binet and the Wechsler

Memory Scale, Jensen's (1971) Memory for Numbers Test, or other related tests. These tests have in common the use of numbers, lack of a study period, and immediate recall and they are considered prototypical Level I tests (Jensen, 1993a). Singer, Andrusiak, Reisdorf, and Black (1992) noted that word familiarity and vocabulary influence memory performance. Thus, in order to minimize confounds and unrelated variance, studies were omitted from the meta-analysis if they used tests that used meaningful word associations, such as a paired associates task, or if they required a study period or had delayed recall.

The second general decision rule concerned the type of individual used by the research study. In order to enhance generalizability of results, studies were only included in the meta-analyses if the participants were healthy, average individuals. Thus, studies such as Davidson, Gibby, McNeil, Segal, and Silverman (1950), which used institutionalized patients, were excluded from the meta-analysis.

In addition to these two general decision rules, Study 1 and Study 2 had unique decision rules.

Study 1. The decision rule unique to Study 1 that affected the inclusion of a research study into the meta-analysis was that only research studies that compared Black and White performance on short-term memory tests and reported sample size, means, and standard deviations for both groups were included in the meta-analysis.

Study 2. The decision rule unique to Study 2 also affected the inclusion of a research study into the meta-analysis. In order to assess the predictive validity of short-term memory tests in applied settings, only studies that used job and training performance criteria were included. These criteria include the traditional methods of measuring performance such as supervisor–peer ratings, quantity of production, work samples, and training course grades. Also, job and training performance criteria would be examined in separate analyses.

Literature Search and Review

Automated and manual searches of books and articles from 1900 to 1994 were conducted. These searches included reviewing the PsychLit databases, bibliographies of relevant, seminal books (e.g., Jensen, 1980; Matarazzo, 1972; Shuey, 1966), the Validity Exchange Index from *Personnel Psychology,* reference lists from the Stanford–Binet (Thorndike, Hagen, & Sattler, 1986) and WAIS intelligence tests, test reviews of the WAIS (e.g., Guertin, Ladd, Frank, Rabin, & Heister, 1966, 1971; Guertin, Rabin, Frank, & Ladd, 1962), and unpublished dissertations and theses. Research studies were included in the meta-analyses based on the previously mentioned decision rules.

The authors imagined that because of the massive amount of research conducted on intelligence testing and ethnicity, there would be many studies that met

the decision rules. Unfortunately, this was not the case. In fact, many studies did not meet the pertinent decision rules for a variety of reasons. Many studies looked at short-term memory differences between SES groups, but not Black and White groups. Several studies used only Black or only White participants, thus no comparison was possible. Other studies failed to report means and standard deviations, failed to report short-term memory scores separately (i.e., only reported full, verbal, and performance scales), failed to report short-term memory test/performance criteria separately (i.e., reported validity for entire test battery only), failed to use statistics amenable to the present analyses (e.g., reported percentiles, percentages, factor loadings), or did not use a short-term memory test that met our decision rules.

For Study 1, the literature search and review yielded 17 studies and 31 Black–White short-term memory test comparisons. These studies included a total of 27,973 individuals. Of the 31 effect sizes, 16 are for adults and 15 are for children. See the Appendix for a list of the studies included in Study 1 and Study 2.

For Study 2, the literature search yielded 11 studies and 141 validity coefficients, covering a total of 34,262 individuals. Of these 141 validity coefficients, 106 are for job performance and 35 are for training performance. Additionally, of the 141 coefficients, 125 came from Ghiselli (1966). Ghiselli (1966) reported the average correlations for two short-term memory tests (immediate recall and substitution) in a variety of occupations categorized by several methods. The validity coefficients associated with Ghiselli's *Dictionary of Occupational Titles* listing were included in the meta-analysis. Because Ghiselli reported average correlations, the sample size was not specific for each coefficient. Instead, Ghiselli reported a range of sample size, such as n = 100–499. In order to use Ghiselli's data, the middle value of the sample size range was used. For example, for the range n = 100–499, the value of 250 was used. The middle value was chosen because using the extreme ends of the range would lead to either an over- or underestimation of the sampling error.

Past research demonstrates that personnel screening tools often have greater validity for some occupations than for others. For example, cognitive ability measures have been found to be more valid for more cognitively demanding jobs than for less demanding jobs (Gutenberg, Arvey, Osburn, & Jeanneret, 1983; Hunter, 1983; McDaniel, 1986b). Psychomotor measures are less valid for more cognitively demanding jobs (Hunter, 1983). For the purpose of moderator analysis, occupations may be classified by job content (e.g., secretaries, police officers), or by job attributes such as the levels of cognitive demands placed on the employee. As discussed by McDaniel, Schmidt, and Hunter (1988a), classification schemes based on job content may typically provide control over sources of variance caused by job attributes because many job attributes have little or no variance within a job family. However, this control is gained at the sacrifice of detailed information about the attributes that may moderate validity. Furthermore, job content classifications require that many validity studies be conducted

for the same job. Job attribute classifications provide a better opportunity to understand why the validity of a given predictor varies across jobs. The success of a job attribute analysis requires the measurement of a large number of jobs on the attributes of interest.

In this study, the job attribute of cognitive demand was measured by attributes provided in the *Dictionary of Occupational Titles* (*DOT;* U.S. Department of Labor, 1977). The authors of the dictionary argued that "every job requires a worker to function to some degree in relation to data, people, and things" (p. xvii). The *DOT* data scale considers synthesizing and coordinating to be high in the data function, whereas copying and comparing jobs are low in the data function. The *DOT* data variable is a measure of the cognitive demands of jobs (Rivkin & McDaniel, 1990). We divided the effect sizes into three categories based on the cognitive demand of the occupation: high (0, 1, and 2), medium (3, 4, and 5), and low (6, 7, and 8). This moderator is also of interest given the findings that short-term memory tests are less predictive for individuals with high general cognitive ability; that is, the individuals who would likely be performing the jobs with high cognitive demands (Jensen, 1993a; Vernon, 1987).

Procedure

In order to meet the goals of this study, two sets of meta-analyses were performed. One set of analyses examined White–Black differences in scores on short-term memory tests. The second set examined the validity of short-term memory tests for predicting job and training performance.

The Hunter–Schmidt psychometric meta-analysis method (Hunter & Schmidt, 1990) used is based on the hypothesis that much of the variation in results across studies may be due to statistical and methodological artifacts rather than to substantive differences in underlying population relationships. Some of these artifacts also reduce the effect sizes (e.g., *r*'s, *d*'s) below their true (e.g., population) values. The method determines the variance attributable to sampling error and to differences between studies in reliability and range restriction, and subtracts that amount from the total amount of variation, yielding estimates of the true variation across studies and of the true average effect size (Hunter & Schmidt, 1990).

To estimate the population distributions of the racial effect sizes and test validities, information on the reliability of short-term memory tests is required, Table 1 presents the set of test–retest reliabilities for short-term memory tests that were located for the present analyses.

In the analysis of the White–Black score differences, a meta-analysis of the standardized mean differences was used. A standardized mean score difference (*d*) expresses the mean score differences in standard deviation units. A *d* value of zero would indicate that there were no mean score differences between Black and White groups. Most cognitive tests yield a *d* of approximately 1, indicating that the mean score of the Black and White groups differs by a full standard deviation. We conducted two sets of analyses. In the first set, the sole artifact correc-

TABLE 1
Test–Retest Reliability Distribution for Short-Term Memory Tests

Test–Retest Reliability	Age/Test Description/Source
.70	Age 16–17; Digit Span; Wechsler (1981)
.89	Age 25–34; Digit span; Wechsler (1981)
.85	Age 35–44; Digit Span; Wechsler (1981)
.82	Age 45–54; Digit Span; Wechsler (1981)
.73	Age 16–17; Digit Symbol; Wechsler (1981)
.86	Age 25–34; Digit Symbol; Wechsler (1981)
.84	Age 35–44; Digit Symbol; Wechsler (1981)
.82	Age 45–54; Digit Symbol; Wechsler (1981)
.72	Adults; Memory for Number Test; Durning (1969)

tion made was for sampling error. This analysis provides the estimated mean race differences that would be observed when using short-term memory tests. These mean observed race differences are underestimates of the population race differences due to the measurement error in the short-term memory tests. The second set of analyses estimate the population race differences by correcting the distribution of observed effect sizes for both sampling error and measurement error in the short-term memory tests. In these analyses, the population mean has been corrected for measurement error in the short-term memory tests. The variance of the population distribution has been corrected for sampling error and differences across studies in the reliability of the short-term memory tests. When the sample consisted of children, the reliability was estimated at .72, which is the mean of the two reliabilities in Table 1 for 16–17-year-olds. Note that the typical mean age of the children in the samples was under age 16. We used the 16–17-year-old reliabilities because they were estimated from the closest age group to our child samples. When the sample consisted of adults, the reliability was estimated at .83, which is the mean of the remaining reliabilities in Table 1.

Studies included in the Black–White meta-analysis used one of four types of short-term memory tests. These tests are: forward digit span, digit symbol substitution, composite tests (i.e., a battery of short-term memory tests such as the short-term memory subtests of the Stanford–Binet), and applied short-term memory. The applied short-term memory tests were developed to select police officers (Barrett & Associates, 1990, 1991, 1992; Barrett, Carobine, & Doverspike, in press). These tests used picture–number paired associates. These tests did not stringently meet the decision rules as they had a brief study period (4 minutes), although they did meet the decision rule of having minimal verbal content. These data were included in the meta-analysis because the tests were developed specifically to reduce adverse impact in an applied setting and thus represent a tailored short-term memory test that differs from more general tests. Black–

White score differences on these four types of short-term memory tests were assessed overall and separately in the analysis.

In the meta-analysis of validity coefficients, we used artifact distribution meta-analysis, using the interactive method (Hunter & Schmidt, 1990, Chapter 4). The mean observed correlation ($\bar{r}$) was used in the sampling error variance formula (Hunter & Schmidt, 1990, pp. 208–210; Law, Schmidt, & Hunter, 1994; Schmidt et al., 1993). The computer program utilized is described in McDaniel (1986a). Additional detail on the program is presented in Appendix B of McDaniel, Schmidt, and Hunter (1988a). The reliability artifact distribution consisted of those reliabilities in Table 1 excluding the two reliabilities for the 16–17-year-olds.

Performance and training criteria were analyzed separately. Scant information was available on the range restriction and criterion reliability for the reported coefficients. Therefore, the job and training performance criterion reliability and range restriction distributions used by Pearlman (1979) were used in this study (average job performance criterion reliability = .60; average training performance criterion reliability = .80). We assert that the use of a job performance criterion reliability distribution with a mean value of .60 is conservative (i.e., underestimates the true validity of the predictors), as Rothstein (1990) found that across 9,975 employees and across all time periods of supervisory exposure to employees, the mean interrater agreement (reliability for one rater) was .48.

RESULTS

Study 1

Table 2 presents the meta-analysis results for the mean score differences on short-term memory tests for Whites and Blacks. The first column of the table identifies the distribution of effect sizes analyzed, and the next two columns present the total number of *d* coefficients on which each distribution was based and the total sample size. The next five columns present the uncorrected mean and standard deviation of each distribution, the standard deviation expected due to sampling error, the percentage of variance due to sampling error, and the residual standard deviation. The residual standard deviation is the estimated standard deviation of the observed distribution corrected for sampling error. The next four columns concern the estimated population distribution. The mean and standard deviation of the estimated population distribution are presented. The last two columns present the 95% credibility interval around the population mean.

Study 2

Table 3 presents the meta-analysis results separately for job and training performance. The first column of the table identifies the distribution of validities analyzed. The next two columns present the number of validity coefficients on which each distribution was based and the total sample size. The mean and stan-

TABLE 2
Meta-Analysis Results for the Mean Score Differences on Short-Term Memory Tests for Whites and Blacks

Distribution	Number of d	Total N	Observed Distribution: Mean d_o	σ_{d_o}	σ_{sc}	% Variance Due to Sampling Error	Residual σ	Population Distribution: Mean d	σ_d	95% Credibility Interval	
All effect sizes	31	27,973	.42	.22	.07	9.3	.21	.48	.24	.01	.95
By Type of Short-Term Memory Measure											
Digit span	16	12,898	.40	.25	.07	8.3	.24	.47	.27	−.07	1.01
Substitution	8	7,521	.35	.19	.07	12.6	.17	.40	.19	.04	.77
Composite battery	3	4,402	.58	.11	.05	25.4	.09	.66	.09	.48	.84
Applied tests	4	3,152	.43	.20	.07	13.4	.18	.48	.20	.08	.87
By Age of Sample											
Adults	16	8,891	.49	.24	.09	13.1	.22	.54	.24	.06	1.02
Children	15	19,082	.39	.20	.06	7.7	.20	.46	.23	.00	.91

Note. Mean d_o = Mean observed d. σ_{d_o} = Standard deviation of the observed d. σ_{sc} = Estimated standard deviation due to sampling error. % Variance due to sampling error = Estimated percentage of observed variance due to sampling error. Residual σ = Estiamted standard deviation of the observed d with sampling error removed. Mean d = Estimated mean population d. σ_d = Estimated standard deviation of population distribution. 95% Credibility Interval = Credibility interval for the population distribution.

TABLE 3
Meta-Analysis Results for the Validity of Short-Term Memory Tests for Job and Training Performance

Distribution	Number of r	Total N	Observed Distribution Mean r	Observed Distribution σ_r	Population Distribution Mean ρ	Population Distribution σ_ρ	90% CV
Job Performance Criteria							
All occupations	106	17,741	.19	.13	.41	.22	.13
Job Performance by Occupation's Level of Cognitive Demands							
High cognitive demands	20	983	.14	.13	.29	.00	.29
Medium cognitive demands	31	6,785	.25	.15	.51	.27	.17
Low cognitive demands	55	10,000	.16	.10	.34	.14	.17
Training Performance Criteria							
All occupations	35	16,521	.28	.08	.49	.09	.38

Note. Mean r = Mean observed correlation coefficient. σ_r = Standard deviation of distribution of observed correlation coefficients. Mean ρ = Estimated mean population correlation coefficient. σ_ρ = Estimated standard deviation of the population distribution. 90% CV = Bottom 10th percentile credibility value.

dard deviation of the distribution of observed coefficients is presented in the next two columns. The final three columns present the estimated population mean (ρ), the estimated population standard deviation (σ_ρ), and the 90% credibility value for the distribution of true validities. This population distribution is corrected for unreliability in the criterion and range restriction. Corrections to the mean do not include corrections for predictor unreliability. The variances of the true validity distributions are corrected for sampling error and for differences among the studies in predictor and criterion reliability and range restriction.

DISCUSSION

Discussions of race differences in test scores typically focus on observed differences that are attenuated by measurement error in the tests. For consistency with this literature, our discussion focuses primarily on the observed distribution mean. The meta-analysis performed in Study 1 revealed that although there are differences between Blacks and Whites on short-term memory tests (d = .42), these differences are less than half the size of those typically found on general cognitive ability measures, which have a d of approximately 1.0 (Gordon, 1986;

Herrnstein & Murray, 1994). Additionally, the smallest group differences were found on the substitution (digit symbol), $d = .35$, and forward digit span tests, $d = .40$. This finding is reasonable, as these types of short-term memory tests are considered prototypical short-term memory measures (Jensen, 1971). Additionally, the applied short-term memory tests resulted in only slightly larger group differences relative to the other tests, $d = .43$.

The composite battery tests resulted in the largest group difference, although the difference is still much less than the difference found on measures of general cognitive ability. However, one should place less confidence in the results for the composite test distribution because it contains only three effect sizes. Distributions with few effect sizes have greater potential for second-order sampling error, which can distort the mean effect size estimate to some extent (Hunter & Schmidt, 1990; Schmidt, Hunter, Pearlman, & Hirsh, 1985, Questions and Answer number 25). Therefore, these analyses should be rerun in the future as more studies become available.

Some samples were composed of adults and others were composed of children. Using the adult versus children dichotomy as a moderator, analyses found that short-term memory differences are more pronounced in adults ($d = .49$) than in children ($d = .39$). This difference is not surprising based on the understanding that intellectual abilities are developmental in nature (Brainerd & Reyna, 1993; Cronbach, 1990; Kamii, 1986). This finding is consistent with those found by Jensen (1971, 1974) in his initial research, which led him to suggest changes in teaching styles and methods in elementary education to prevent greater differences at later ages.

An inspection of Table 1 reveals that short-term memory tests have lower reliabilities for children than adults. Thus, one might speculate that the difference in race effect sizes between children and adults might be an artifact of the reliability of the tests in the two populations. An inspection of the population distribution means indicate that when the effect sizes are corrected for test measurement error, the mean effect size for adults (.54) is still larger than the effect size for the children (.46). Thus, the effect size differences in observed test scores are not primarily due to test reliability differences.

In summary for Study 1, the large sample of 27,973 individuals analyzed permits one to place great confidence in the finding that tests of short-term memory result in smaller racial mean differences than tests of general cognitive ability. Although the degree of race differences varies somewhat by type of short-term memory test and by the age of the sample, all distributions show substantially smaller race differences then do tests of general cognitive ability.

The meta-analyses performed in Study 2 revealed that short-term memory tests are valid predictors of both job performance ($\rho = .41$) and training performance ($\rho = .49$). We offer a post hoc explanation for the curvilinear relationship by complexity found in the job performance data. High-complexity job, such as

executive or manager, are likely to have a large general cognitive ability component and minimal short-term memory component, thus short-term memory tests should have a lower correlation with performance on these jobs. Similarly, jobs with a medium cognitive complexity level, such as secretary and general clerk, should have a relatively high short-term memory component based on the types of tasks performed. Low cognitive complexity jobs, such as produce packer, are often largely based on manual labor, and thus have a smaller emphasis on short-term memory abilities. This pattern suggests that short term memory tests are valid predictors for all job levels, but predict some jobs more strongly than others.

Study 2 also found that short-term memory tests predict training performance ($\rho = .49$) better than job performance ($\rho = .41$). This pattern of validities is consistent with the research on the validity of cognitive ability measures. In general, cognitive ability measures yield higher validities for training than for performance criteria (Lilienthal & Pearlman, 1983; Pearlman, 1979).

Validity may be concluded to be generalizable if the value at the lower 10th percentile of the distribution of estimated true validities is greater than zero (Callender & Osburn, 1981). This definition of validity generalizability is directly analogous to significance testing. A correlation is statistically significant when the lower bound of its confidence interval is above zero. By this criterion, all distributions in Table 2 show validity generalization.

The meta-analyses conducted in Study 1 and Study 2 demonstrate that short-term memory tests are valid predictors of job and training performance and result in less adverse impact than tests of general cognitive ability. This finding identifies a method that assists employers in meeting affirmative action goals and legal mandates. Additionally, such findings give both researchers and practitioners hope that applied duties can be met without compromising professional integrity.

As noted by reviewers of this manuscript, our results are consistent with theory and research on *g*. Jensen (1985), among others, theorized that the Black–White difference on test scores is a function of the extent to which the tests tap *g*. Whereas short-term memory measures are less related to *g* than general cognitive ability measures, an advocate of Jensen's (1985) position would expect short-term memory tests to yield smaller race differences than general cognitive measures. There is also substantial evidence that the validity of any cognitive test is primarily due to the extent to which it taps *g* (Olea & Ree, 1994; Ree, Earles, & Teachout, 1994; Thorndike, 1986). Although the validity of measures of general cognitive ability varies by job, a validity coefficient of .50 is typical for most measures of general cognitive ability (Hunter & Hunter, 1984). When compared to the .41 validity of short-term memory measures for job performance, one may conclude that the reduced race differences are purchased at the cost of reduced validity. The critical question that employers then face is how much validity reduction one is willing to accept for the reduction in race differences. Given that many employers routinely compromise strict merit hiring to promote a racially

balanced workforce, we suspect that many will find a test with a relatively high validity of .41 and relatively low adverse impact to be very acceptable.

The search for predictors with high validity and low racial differences has been the search for the Holy Grail in personnel psychology. Like the search for the grail, most attempts have met with failure. Thus, the results reported here, which offer the promise of high validity and low racial differences, should be examined most carefully. Here, we offer two considerations and potential limitations of the presented research.

First, we note that variance of the *d* effect sizes is substantial. Some studies show more race differences than others and this variability is not totally explained by sampling error and differences across studies in the reliability of the short-term memory tests. Most of this variance cannot be attributed to differences in the types of short-term memory tests. When we subdivided the full data set into four categories of short-term memory measures, the effect sizes varied somewhat by type of test, but within each test category substantial variance remained. Further disaggregation of the types of short-term memory tests cannot be justified given the low number of effect sizes in our current data set. Furthermore, based on the content similarity of the various types of short-term memory tests, we do not believe that the type of short-term memory measure will prove to a be a substantial source of variance in the distribution.

In addition, the variance in the racial effect sizes cannot be fully attributed to whether the participants in the analysis were adults or children. When we subdivided the full data set into adult and child samples, the adult samples showed larger race differences than the child samples, yet each of these subdistributions still showed large variance.

Although we offer no well-researched hypothesis to explain the variability of effect sizes, we speculate that one cause of this variability is that some samples are more variable than others on nonrace variables related to short-term memory. For example, samples in which both the Blacks and Whites are drawn from the same socioeconomic strata (e.g., samples from wealthy suburbs) might show smaller race differences than samples including individuals from a variety of socioeconomic strata.

Second, we suggest caution in interpreting the moderating effects of cognitive demands on the validity of the short-term memory tests. For the distribution of validity coefficients for high-complexity jobs, the number of coefficients (20) and the number of individuals contributing to the studies (983), although not minuscule, is not as numerous as we would prefer. We suggest that the relatively low validities for these high-complexity occupations may be due to second-order sampling error.

In conclusion, we offer that our findings of high validity with low adverse impact have substantial import for the practice of personnel selection. When short-term memory tests are optimally weighted in a selection composite, the resulting composite will have higher validity and lower adverse impact than the

composite without the short-term memory test. We are aware that these results may be surprising to most and noncredible to some. Both the potential substantial import of the results and the limitations of this study suggest that personnel psychologists should devote much more attention to the potential role of short-term memory tests in personnel selection.

APPENDIX: STUDIES INCLUDED IN META-ANALYSES

The following are lists of the actual articles included in the meta-analyses for both Study 1 and Study 2.

Study 1

Barrett & Associates, Inc. (1990). *Technical report: Midwestern city police entrance exam.* Akron, OH: Author.

Barrett & Associates, Inc. (1991). *Technical report: Midwestern city police entrance exam.* Akron, OH: Author.

Barrett & Associates Inc. (1992). *Technical report: Midwestern city firefighter entrance exam.* Akron, OH: Author.

Barrett, G.V., Carobine, R.G., & Doverspike, D. (in press). The reduction of adverse impact in an employment setting using a short-term memory test. *Journal of Business and Psychology.*

Buckhalt, J.A., Denes, G.E., & Stratton, S.P. (1989). Validity of the British Ability Scales Short-Form for a sample of U.S. students. *School Psychology International, 10,* 185–191.

Durning, K.P. (1969). *Preliminary assessment of the Navy Memory for Numbers test.* Unpublished master's thesis, San Diego State College, San Diego, CA.

Grandison, F.L. (1951). *The relationship of level of aspiration to negro and white differences on form I of the Wechsler-Bellevue Intelligence Scale.* Unpublished master's thesis, The Ohio State University, Columbus, OH.

Hall, V.C., & Kaye, D.B. (1980). Early patterns of cognitive development. *Monographs of the Society for Research in Child Development, 45,* 1–74.

Jensen, A.R. (1971). Do schools cheat minority children? *Educational Research, 114,* 3–28.

Jensen, A.R. (1974). Interaction of level I and level II abilities with race and socioeconomic status. *Journal of Educational Psychology, 66,* 99–111.

Jensen, A.R., & Figueroa, R.A. (1975). Forward and backward digit span interaction with race and IQ: Predictions from Jensen's theory. *Journal of Educational Psychology, 67,* 882–893.

Jensen, A.R., & Reynolds, C.R. (1982). Race, social class and ability patterns on the WISC-R. *Personality and Individual Differences, 3,* 423–438.

Kaufman, A.S., McLean, J.E., & Reynolds, C.R. (1988). Sex, race, residence, region, and education differences on the 11 WAIS-R subtests. *Journal of Clinical Psychology, 44,* 231–248.

Nichols, P.L. (1970). *The effects of heredity and environment on intelligence test performance in 4 and 7 year old white and negro sibling pairs.* Unpublished doctoral dissertation, University of Minnesota, Minneapolis.

Solkoff, N. (1974). Race of examiner and performance on the Wechsler Intelligence Scale for Children: A replication. *Perceptual and Motor Skills, 39,* 1063–1066.

Thorndike, R.L., Hagan, E.P., & Sattler, J.M. (1986). *Stanford–Binet test manual* (4th ed.). Chicago: Riverside Publishing Company.

Tuttle, L.E. (1964). *The comparative effect on intelligence test scores of negro and white children when certain verbal and time factors are varied.* Unpublished doctoral dissertation, University of Florida, Gainesville.

Study 2

Balinsky, B., & Shaw, H.W. (1956). The contribution of the WAIS to a management appraisal program. *Personnel Psychology, 9*, 207–209.

Dubois, P.H., & Watson, R.I. (1950). The selection of patrolmen. *Journal of Applied Psychology, 34*, 90–95.

Durning, K.P. (1969). *Preliminary assessment of the Navy Memory for Numbers test.* Unpublished master's thesis, San Diego State College, San Diego, CA.

Freyd, M. (1927). Selection of typists and stenographers: Information on available tests. *Journal of Personnel Research, 5*, 490–510. (Includes studies by Bieneman, 1923; Lahy, 1913; and Muscio and Swoton, 1923).

Ghiselli, E.E. (1966). *The validity of occupational aptitude tests.* New York: Wiley.

Meer, B., & Stein, M.I. (1955). Measures of intelligence and creativity. *Journal of Psychology, 39*, 117–126.

Skula, M., & Spillane, R.F. (1954a). Validity information exchange. *Personnel Psychology, 7*, 136–137.

Skula, M., & Spillane, R.F. (1954b). Validity information exchange. *Personnel Psychology, 7*, 147–148.

Voicu, C., & Nereuta, A.L. (1985). Relation between aptitude level and professional efficiency with operators who control complex supervising and command equipment involving failure and accident producing risks. *Revue Roumaine des Sciences Sociales: Serie de Psychologie, 29*, 131–137.

REFERENCES

Ash, R.A., Johnson, J.C., Levine, E.L., & McDaniel, M.A. (1989). Job applicants training and work experience evaluation in personnel selection. In G.R. Ferris & K.M. Rowland (Eds.), *Research in Personnel and Human Resources Management, 7*, 183–226. Greenwich, CT: JAI Press.

Barrett & Associates, Inc. (1990). *Technical report: Midwestern city police entrance exam.* Akron, OH.

Barrett & Associates, Inc. (1991). *Technical report: Midwestern city police entrance exam.* Akron, OH.

Barrett & Associates Inc. (1992). *Technical report: Midwestern city firefighter entrance exam.* Akron, OH.

Barrett, G.V., Carobine, R.G., & Doverspike, D. (1992). *The reduction of adverse impact in an employment setting using a short-term memory test.* Manuscript submitted for publication.

Barrick, M.R., & Mount, M.K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*, 1–26.

Brainerd, C.J., & Reyna, V.F. (1993). Memory independence and memory interference in cognitive development. *Psychological Review, 100*, 42–67.

Callender, J.C., & Osburn, H.G. (1981). Testing the constancy of validity with computer-generated sampling distributions of the multiplicative model variance estimate: Results for the petroleum industry validation research. *Journal of Applied Psychology, 66*, 274–281.

Cronbach, L.J. (1990). *Essentials of psychological testing.* New York: Harper & Row.

Davidson, K.S., Gibby, R.G., McNeil, E.B., Segal, S.J., & Silverman, H. (1950). A preliminary study of negro and white differences on form I of the Wechsler-Bellevue scale. *Journal of Consulting Psychology, 14*, 489–492.

Dye, D.A., & Reck, M. (1989). College grade point as a predictor of adult success: A reply. *Public Personnel Management, 18*, 235–241.

Ghiselli, E.E. (1966). *The validity of occupational aptitude tests.* New York: Wiley.

Gordon, R.A. (1986). Scientific justification and the race-IQ-delinquency model. In T.F. Hartnagel & R.A. Silverman (Eds.), *Critique and explanation: Essays in honor of Gwynne Nettler*. New Brunswick, NJ: Transaction Books.

Guertin, W.H., Ladd, C.E., Frank, G.H., Rabin, A.I., & Heister, D.S. (1966). Research with the Wechsler intelligence scales for adults: 1960–1965. *Psychological Bulletin, 66*, 385–409.

Guertin, W.H., Ladd, C.E., Frank, G.H., Rabin, A.I., & Heister, D.S. (1971). Research with the Wechsler intelligence scales for adults: 1965–1970. *The Psychological Record, 21*, 289–339.

Guertin, W.H., Rabin, A.I., Frank, G.H., & Ladd, C.E. (1962). Research with the Wechsler intelligence scales for adults: 1955–1960. *Psychological Bulletin, 59*, 1–26.

Gutenberg, R.L., Arvey, R.D., Osburn, H.G., & Jeanneret, P.R. (1983). Moderating effects of decision-making/information-processing job dimensions on test validities. *Journal of Applied Psychology, 54*, 27–30.

Herrnstein, R.J., & Murray, C.A. (1994). *The bell curve: Intelligence and class structure in American life*. New York: The Free Press.

Hunter, J.E. (1983). *Test validation for 12,000 jobs: An application of job classification and validity generalization analysis to the General Aptitude Test Battery (GATB)*. Washington, DC: Division of Counseling and Test Development, Employment and Training Administration, U.S. Department of Labor.

Hunter, J.E., & Hunter, R.F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 98*, 72–98.

Hunter, J.E., & Schmidt, F.L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Beverly Hills, CA: Sage.

Jensen, A.R. (1971). Do schools cheat minority children? *Educational Research, 114*, 3–28.

Jensen, A.R. (1974). Interaction of level I and level II abilities with race and socioeconomic status. *Journal of Educational Psychology, 66*, 99–111.

Jensen, A.R. (1980). *Bias in mental testing*. New York: MacMillan.

Jensen, A.R. (1985). The nature of black-white differences on various psychometric tests: Spearman's hypothesis. *The Behavioral and Brain Sciences, 8*, 193–264.

Jensen, A.R. (1993a). Psychometric *g* and achievement. In B.R. Gifford (Ed.), *Policy perspectives on educational testing*. Boston: Kluwer.

Jensen, A.R. (1993b). Spearman's hypothesis tested with chronometric information-processing tasks. *Intelligence, 17*, 44–77.

Jensen, A.R., & Figueroa, R.A. (1975). Forward and backward digit span interaction with race and IQ: Predictions from Jensen's theory. *Journal of Educational Psychology, 67*, 882–893.

Kamii, C. (1982). *Number in preschool and kindergarten*. Washington, DC: National Association for the Education of Young Children.

Law, K.S., Schmidt, F.L., & Hunter, J.E. (1994). Non-linearity of range correlations in meta-analysis: A test of an improved procedure. *Journal of Applied Psychology, 79*, 425–438.

Lilienthal, R.A., & Pearlman, K. (1983). *The validity of federal selection tests for aid/technicians in the health, science, and engineering fields*. Washington, DC: U.S. Office of Personnel Management, Office of Personnel Research and Development.

Matarazzo, J.D. (1972). *Wechsler's measurement and appraisal of adult intelligence*. Baltimore: Williams & Wilkins.

McDaniel, M.A. (1986a). Computer programs for calculating meta-analysis statistics. *Educational and Psychological Measurement, 46*, 175–177.

McDaniel, M.A. (1986b). The evaluation of a causal model of job performance: The interrelationships of general mental ability, job experience, and job performance. (Doctoral dissertation, The George Washington University). *Dissertation Abstracts International, 46*, AAD86–08356.

McDaniel, M.A., Schmidt, F.L., & Hunter, J.E. (1988a). A meta-analysis of methods for rating training and experience in personnel selection. *Personnel Psychology, 41,* 283–314.

McDaniel, M.A., Schmidt, F.L., & Hunter, J.E. (1988b). Job experience correlates of job performance. *Journal of Applied Psychology, 73,* 327–330.

McDaniel, M.A., Whetzel, D., Schmidt, F.L., & Maurer, S. (1994). The validity of the employment interview: A comprehensive review and meta-analysis. *Journal of Applied Psychology, 79,* 599–616.

Olea, M.M., & Ree, M.J. (1994). Predicting pilot and navigator criteria: Not much more than g. *Journal of Applied Psychology, 79,* 845–851.

Ones, D.S., Viswesvaran, C., & Schmidt, F.L. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance. *Journal of Applied Psychology, 78,* 679–703.

Pearlman, K. (1979). *The validity of tests used to select clerical personnel: A comprehensive summary and evaluation (TS–79–1)* (NTIS No. PB 80. 102 650). Washington, DC: U.S. Office of Personnel Management, Personnel Research and Development Center.

Raven, J.C., Court, J.H., & Raven, J. (1977). *Manual for Raven's Progressive Matrices and Vocabulary scales.* London: H.K. Lewis & Co.

Ree, M.J., Earles, J.A. & Teachout, M.S. (1994). Predicting job performance: Not much more than *g. Journal of Applied Psychology, 79,* 518–524.

Rivkin, D., & McDaniel, M.A. (1990, August). The measurement and validation of occupational aptitude requirements. In A. Lancaster (Chair), *The enhancement of the department of defense student testing program.* Symposium presented at the 98th Annual conference of the American Psychological Association, Boston.

Rothstein, H.R. (1990). Interrater reliability of job performance ratings: Growth to asymptote level with increasing opportunity to observe. *Journal of Applied Psychology, 75,* 322–327.

Rothstein, H.R., & McDaniel, M.A. (1989). Guidelines for conducting and reporting meta-analyses. *Psychological Reports, 65,* 759–770.

Rothstein, H.R., Schmidt, F.L., Erwin, F.W., Owens, W.A., & Sparks, C.P. (1990). Biographical data in employment selection: Can validities be made generalizable? *Journal of Applied Psychology, 75,* 175–184.

Sattler, J. (1988). *Assessment of children's intelligence and other special abilities* (2nd ed.). Boston: Allyn & Bacon.

Schmidt, F.L., Hunter, J.E., Pearlman, K., & Hirsh, H.R. (1985). Forty questions and answers about validity generalization and meta-analysis. *Personnel Psychology, 38,* 697–798.

Schmidt, F.L., Law, K., Hunter, J.E., Rothstein, H.R., Pearlman, K., & McDaniel, M.A. (1993). Refinements in validity generalization procedures: Implications for the situational specificity hypothesis. *Journal of Applied Psychology, 78,* 3–13.

Shuey, A.M. (1966). *The testing of Negro intelligence.* New York: Social Science Press.

Singer, M., Andrusiak, P., Reisdorf, P., & Black, N.L. (1992). Individual differences in bridging inference processes. *Memory and Cognition, 20,* 539–548.

Thorndike, R.L. (1986). The role of general ability in prediction. *Journal of Vocational Behavior, 29,* 332–339.

Thorndike, R.L., Hagan, E.P., & Sattler, J.M. (1986). *Stanford–Binet test manual.* Chicago: Riverside Publishing Company.

Vernon, P.A. (1987). Part III: Human learning: Level I/II theory. In S. Modgil & C. Modgil (Eds.), *Arthur Jensen: Consensus and controversy.* New York: Falmer.

Vincent, K.R. (1991). Black/white IQ differences: Does age make the difference? *Journal of Clinical Psychology, 47,* 266–270.

Wechsler, D. (1958). *The measurement and appraisal of adult intelligence.* Baltimore: Williams & Wilkins.

Wechsler, D. (1981). *Wechsler Adult Intelligence Scale—Revised manual.* San Antonio, TX: The Psychological Corporation.

Wonderlic, E.F., & Wonderlic, C.F. (1972). *Wonderlic Personnel Test: Negro norms.* Northfield, IL: E.F. Wonderlic & Associates.

U.S. Department of Labor. (1977). *Dictionary of occupational titles* (4th ed.). Washington, DC: Employment and Training Administration, U.S. Employment Service.