Psychology, Public Policy, and Law
1996, Vol. 2, No. 3/4, 418–446

1076-8971/96/$3.00

RACIALLY GERRYMANDERING THE CONTENT OF POLICE TESTS TO SATISFY THE U.S. JUSTICE DEPARTMENT: A Case Study

Linda S. Gottfredson
University of Delaware

Discrimination law and its aggressive enforcement by the U.S. Department of Justice both falsely assume that all racial–ethnic groups would pass job-related, unbiased employment tests at the same rate. Unreasonable law and enforcement create pressure for personnel psychologists to violate professional principles and lower the merit relatedness of tests in the service of race-based goals. This article illustrates such a case by describing how the content of a police entrance examination in Nassau County, New York, was stripped of crucial cognitive demands to change the racial composition of the applicants who seemed most qualified. The test was thereby rendered nearly worthless for actually making such determinations. The article concludes by examining the implications of the case for policing in Nassau County, Congressional oversight of Justice Department activities, and psychology's role in helping its members to avoid such coercion.

The influence of politics and government on science has long been a concern in both science and society. I focus here on one aspect of that influence as it relates to psychology. What are the responsibilities of psychologists when federal law or its enforcement agencies press them to implement scientific theories that have been proven false or to violate their professional standards for political ends? Who bears responsibility if harm results from their acceding to such government pressure, especially without the client's knowledge? Also, how can psychology protect its members against such coercion in the first place?

I do not have the answers to these questions. However, the following case study illustrates that the failure to address them harms both psychology and the society that law is meant to protect. I begin with an abbreviated account of events surrounding the development of a police entrance examination in Nassau County, New York, and then describe (a) the false assumption that the U.S. Department of Justice expects psychologists in such settings to implement and the professional dilemma it creates, (b) the various means by which personnel psychologists effect compliance with the false assumption, and (c) how compliance was achieved with the new Nassau County test. I conclude by looking at the implications of the new exam for the quality of policing in Nassau County, the questions Congress might ask about the Department of Justice's distorted enforcement of already unreasonable laws and regulations, and the ethical guidelines psychology might provide its practitioners when enforcement agencies pursue objectives that are inconsistent with their profession's established standards and even support their violation.

It should be noted in fairness that the general path of compliance that I describe has been well trodden in personnel selection during the last two decades.

Correspondence concerning this article should be addressed to Linda S. Gottfredson, Department of Educational Studies, University of Delaware, Newark, Delaware 19716. Electronic mail may be sent via Internet to gottfred@udel.edu.

The Nassau County case stands out primarily for the skill and knowledge of the individuals involved, their unprecedented partnership with the Justice Department, and the national ramifications of that relationship.

A Short History

The Promise

During 3 days in 1994, over 25,000 people took Nassau County's new police entrance examination: *Nassau County [NY] Police Officer Examination No. 4200.* In late 1996, the county selected its first training class of under 100 recruits. During the next few years, the county expects to screen the top 20% of scorers on the test and actually hire no more than about 3% of the applicants.

In July 1995 an illustrious team of industrial psychologists released a technical report detailing their "innovative" procedures in developing the exam, which they said would "improve on 'typical selection practices' " (HRStrategies, 1995, p. 12). It thus appeared that the Nassau County Police Department was in an enviable position as an employer. With both a large pool of applicants and what was promised to be an effective tool for identifying the very best among them, the department could improve its already highly professional corps of police officers.

The Nassau County Police Department had been sued by the U.S. Department of Justice in 1977 for employment discrimination, and its subsequent recruiting and hiring were governed by a long series of consent decrees. The 1994 exam had been developed pursuant to a 1990 consent decree. That decree specified that Nassau County and the Justice Department agreed to jointly "develop a new exam that either does not have adverse impact upon blacks, Hispanics and females, or has been validated [shown to be job-related]" (*United States v. Nassau County,* 1995a, p. 2). The new test's 1983 and 1987 predecessors, also developed under consent decrees, had both been litigated because they had substantial disparate impact. In contrast, the 1994 exam had no disparate impact on Hispanics and women and relatively little on Blacks. It therefore seemed to promise that the county could finally end two decades of litigation.

The special counsels for both the county and the Justice Department lauded the test in seeking approval for its use from the U.S. District Court. William H. Pauley, III, the county's special counsel to the police department over the many years of Justice Department litigation, stated that

> the 1994 Examination is now recognized by DoJ [Department of Justice] and industrial psychologists as the finest selection instrument for police officers in the United States. (*Hayden et al. v. Nassau County,* 1996a, pp. 15–16)

John M. Gadzichowski, Justice's representative in the 1977 suit and in subsequent consent decrees, testified that "it's beyond question that the examination . . . is valid" and that "it's the closest ['to a perfect exam, vis-à-vis the adverse impact'] that I've seen in my years of practice" (*United States v. Nassau County,* 1995b, pp. 22–24, 26).

Soon after the new exam received the District Court's approval in fall 1995, the Justice Department began encouraging other police departments around the nation to consider adopting some version of the Nassau test. Aon Consulting, the consulting firm that had led development of the test (at that time named HRStrategies), simultaneously issued a widely circulated invitation in spring 1996

(Aon Consulting, 1996b) urging other police departments to join a test validation consortium. It stated that the project's objective "is to produce yet additional refinements to the Nassau County-specific test, and to reduce even further the level of adverse impact among minority candidates" (Aon Consulting, 1996b, p. 6). The announcement concluded by stressing the legal advantages of joining the consortium: "Ongoing review of the project by Department of Justice experts will provide a device that satisfies federal law" (Aon Consulting, 1996b, p. 7).

Justice's role in this venture clearly suggests that there is legal risk for other police departments if they choose not to try out a Nassau-like test. Under civil rights laws and regulations, when two selection devices serve an employer's needs equally well, the employer must use the one that screens out fewer protected minorities. The Justice Department soon began to treat the Nassau exam as a model for valid, minimally impactful alternatives for police selection. Police departments that failed to switch to Nassau-like tests thus took the risk of being litigated as discriminatory.

Indeed, just months after the court approved the Nassau County test, the National Association for the Advancement of Colored People (NAACP) threatened to sue the New Jersey State Police for discrimination but suggested that litigation might be prevented if the State Police considered switching to the Nassau County test (letter from Joshua Rose, of the law firm representing the NAACP, to Katrina Wright, New Jersey Deputy Attorney General, February 5, 1996, p. 2). Although the test the New Jersey State Police currently uses had itself been developed and adopted several years earlier under pressure from the Justice Department, then represented by David Rose (father and now law partner of Joshua Rose), it screened out more minority applicants than did the Nassau test. The New Jersey State Police refused to change its test and was sued on June 24, 1996 (*NAACP v. State of New Jersey,* 1996).

The jointly developed Nassau County test was an instance in which psychologists worked closely with Justice Department representatives to develop an entrance exam that would be as valid as but have less disparate impact than previous tests in Nassau County. As Justice Department Special Litigation Counsel Gadzichowski explained to the court:

> [M]y department made a decision to break ground. . . . We thought that rather than coming in and challenging an exam every two and three years, so to speak, knocking it out, then coming back three years hence to look at another exam, we would participate in a joint test development project. (*United States v. Nassau County,* 1995b, p. 20)

The Reality

However, the Nassau County test was not what the county or the court was told it was.

The first sign of discontent was local. It came immediately after the July 30 and 31, 1994, administrations of the new exam. There were complaints in local newspapers of inadequate proctoring and rampant cheating during the exam (e.g., Nelson & Shin, 1994), and later it would be reported that more than 40 applicants had been disqualified for cheating (Topping, 1995). The project's "creative [video] examination format" had required that the test be given in Madison Square Garden and the Nassau Coliseum, which posed far greater security problems than the small rooms in which such tests are usually administered.

The next sign of discontent emerged 1 year later when applicants received their test scores. Eighty-five White and Hispanic test takers, half of whom were the sons and daughters of police officers, filed a lawsuit alleging reverse discrimination in the test's development and scoring (*Hayden et al. v. Nassau County,* 1996b). Their suit had been stimulated by what seemed to them to be obvious peculiarities in who received high versus low scores. All the plaintiffs had done very poorly or failed the test despite usually doing well on such tests, yet many others who scored well had a history of poor performance.

The plaintiffs' suspicions about the test had been buttressed by reports leaking out of the police department's background investigation unit. Those reports, from officers afraid to go public, claimed that while some of the top scorers called in for further processing seemed quite good, a surprising number of others were semiliterate, had outstanding arrest warrants, declined further processing when asked to take the drug test, or could not account for years of their adult life. Those who had drug problems, previous convictions, or questionable results on the newly instituted polygraph test would most likely be weeded out. However, the unprecedented poor quality of the candidates who scored well on the new test strongly suggested that something was amiss with the test.

The Justice Department routinely denies that it promotes any particular test or test developer, but it has a history of doing just that (e.g., see O'Connell & O'Connell, 1988, on how the Department of Justice pressured the city of Las Vegas to use the firm of Richardson, Bellows, and Henry [RBH]). As reported by RBH President Frank Erwin, Justice also has a history of trying to coerce its favored developers into, among other things, giving less weight to the cognitive portions of their exams than warranted (personal communication on how RBH's unwillingness to accommodate inappropriate Justice Department requests ended that relation). With Justice's promoting the Nassau exam, members of the professional test development community became increasingly concerned about its interference in test development. To verify their concerns, some of them called on selected academics in June 1996 to evaluate the long technical report describing Nassau County's new test.

I was one of the academics called. We all read the report independently of one another, without prior knowledge of who the test developers were, without prior information about the report's contents or origins, and without compensation offered, expected, or received. (I have never had any financial interest in any testing enterprise.) After reading the report, I obtained court records and interviewed a variety of people in Nassau County and test developers nationwide. In the following months three researchers wrote critiques of the new test (Gottfredson, 1996a, 1996b, 1996c; C. J. Russell, 1996; Schmidt, 1996a, 1996b).

Those evaluations were all highly critical of the report and the test it described. The unanimous opinion was that the concern for hiring more protected minorities had overridden any concern with measuring essential skills. As explained below, the new test may be at best only marginally better than tossing a coin to select police officers—which would explain the mix of both good and bad candidates among the top scorers.

The most distinctive thing about the test is what it omitted—virtually any measurement of cognitive (mental) skills. Although the project's careful job analysis had shown that "reasoning, judgment, and inferential thinking" were the

most critical skills for good police work, the final implementation version of the exam (the one used to rank applicants) retained only personality (*non*cognitive) scales such as Achievement Motivation, Openness to Experience, and Emotional Stability. The reading component of the "experimental" test battery (the version actually administered to applicants the year before) was regraded as pass–fail; to pass that test, applicants only had to read as well as the worst 1% of readers in the research sample of incumbent police officers. Nor did failing the reading component disqualify an applicant because the final exam score was determined by combining the scores from all nine tests. Not mincing words, Schmidt (1996a, 1996b) predicted that the test would be "a disaster" for any police force that used it.

The three commentators' suspicion that the test had been shaped more by Justice's expectation than professional considerations was confirmed by one of Aon's own vice presidents (quoted in Zelnick, 1996):

> Through 18 years and four presidents the message from the Justice Department was clearly that there was no way in Hell they would ever sign onto an exam that had an adverse impact on blacks and Hispanics. What we finally came up with was more than satisfactory if you assume a cop will never have to write a coherent sentence or interpret what someone else has written. But I don't think anyone who lives in Washington [DC] could ever make that assumption. (pp. 110–111)

In referring to the aftermath of Washington, DC's many years of lax hiring, Aon's representative was echoing Schmidt's (1996a, 1996b) prediction of disaster for Nassau County. Among other problems, Washington, DC had developed a "notorious record for seeing felony charges dismissed because of police incompetence in filling out arrest reports and related records" (Zelnick, 1996, p. 111).

The Testing Dilemma

The Justice Department's expectation, like employment discrimination law and regulation in general, is rooted in a false assumption: But for discrimination, all race and gender groups would score equally well on job-related, unbiased employment tests.

This presumption undergrids perhaps the most important element of employment discrimination law and regulation—*disparate impact theory* (Sharf, 1988). Disparate impact theory holds that an employer's failure to hire approximately equal proportions of all races and genders constitutes prima facie evidence of unlawful employment discrimination. The employer then bears the burden of demonstrating that the selection procedure in question is "job related" (merit related) or justified by "business necessity." If the employer succeeds, the burden then shifts to the plaintiffs, who prevail against the employer if they show that there is an alternative selection device that would meet the employer's needs equally well but have less disparate impact.

Disparate impact theory was introduced by two federal regulatory agencies in the late 1960s (see Sharf, 1988, for a history), was incorporated into case law by the Supreme Court's 1971 decision in *Griggs v. Duke Power Co.*, and was made part of statutory law by the Civil Rights Act of 1991. The ways in which regulatory agencies interpret disparate impact law and how the Justice Department enforces it are crucial because these agencies can effectively ban all merit-related (valid) tests

with disparate impact by making it difficult and costly to demonstrate job relatedness to those agencies' satisfaction. This has, in fact, been the game: Drive employers away from valid tests with disparate impact by making it too costly to defend them. A key tool in this game has been the federal government's onerous and scientifically outmoded set of rules for showing the job relatedness of tests, the *Uniform Guidelines for Employee Selection Procedures* (Equal Employment Opportunity Commission [EEOC], Civil Service Commission, Department of Labor, & Department of Justice, 1978).

Since the late 1960s, personnel psychologists have tried to help employers meet the dictates of disparate impact theory and its often unreasonable enforcement. They have become more successful in helping larger (wealthier) organizations to defend merit-related selection procedures in litigation, but their greatest efforts have gone into seeking good procedures that will not trigger litigation in the first place; that is, highly valid tests with little or no disparate impact. These efforts at finding highly merit-related tests with little impact have not been as fruitful as the psychologists had expected and hoped.

Research in the last two decades helps to explain why. The research has provided a fairly clear picture of what kinds of worker traits and aptitudes predict different aspects of job performance and how those traits differ across demographic subgroups (e.g., see the review by T. L. Russell, Reynolds, & Campbell, 1994). It has thus been able to explain why some selection devices have more validity or disparate impact than others and has begun to chart how much of both different selection batteries produce.

The major legal dilemma in selection is that the best overall predictors of job performance (viz., cognitive tests) have the most disparate impact on racial–ethnic minorities. Their considerable disparate impact is not due to any imperfections in the tests. Rather, it is due to the tests' measuring essential skills and abilities that happen not to be distributed equally among groups (Schmidt, 1988). Those differences currently are large enough to cause a major problem. U.S. Department of Education literacy surveys show, for example, that Black college graduates, on the average, exhibit the cognitive skill levels of White high school graduates without any college (Kirsch, Jungeblut, Jenkins, & Kolstad, 1993, p. 127).

This dilemma means that the disparate impact of cognitive tests can only be reduced by diminishing their ability to predict job performance. In fact, that problem is so well-known among personnel selection professionals that there is considerable research estimating how much productivity is lost by lessening the validity of cognitive tests by different degrees to reduce their disparate impact (e.g., Hartigan & Wigdor, 1989; Hunter, Schmidt, & Rauschenberger, 1984; Wigdor & Hartigan, 1988; see also Brody, 1996, for a more general discussion of the same dilemma). There are two general methods of reducing the disparate impact of cognitive tests: lower the hiring standards only for the lower scoring groups or lower standards for all races and ethnicities. Double standards reduce productivity less than low common standards do because they maintain standards for the majority of workers. The drawbacks of double standards are that they are obviously race conscious and that they create disparate impact in future promotions. In contrast, low common standards have the virtue of being race neutral, but they devastate workforce performance across the board.

Unfortunately, current racial disparities in skills and abilities are such that

disparate impact can routinely be expected, at least for Blacks, under race-neutral hiring in most jobs. Moreover, the disparate impact to be expected (and the levels actually found) worsens with the complexity level of the occupation in question (Gottfredson, 1986).

Litigation is very costly, so many employers, particularly in the public sector, prefer to settle out of court or sign consent decrees rather than fight an adverse impact lawsuit. Moreover, as has been observed in many police and fire departments over the last two decades, employers who resist are often litigated by the Justice Department or civil rights groups until they eliminate the disparate impact by whatever means.

Ways of Limiting the Disparate Impact of Cognitive Tests

Showing the merit relatedness of tests with disparate impact, as the law requires, is a straightforward technical matter if the employer's purse is ample enough. Complying with unreasonable enforcement policy is not so simple, however. The Justice Department has been averse to accepting job relatedness data for tests with substantial disparate impact. In technical terms, Justice is effectively requiring employers and their selection psychologists to artificially limit or reduce the validity of many of their selection devices. Whether explicit or covert, witting or not, some psychologists have developed a variety of strategies for doing so.

There are times, of course, when considerations of cost or feasibility prevent employers from using what they know would be better systems for identifying the most capable job candidates. However, job relatedness is often intentionally reduced or limited solely to reduce disparate impact. There are three general ways of doing so with cognitive tests. The first and third decrease job relatedness, whereas the second increases it.

Use Double Standards

Race-norming, or within-group scoring, is the most technically sophisticated method for instituting double standards. It adjusts test scores by race (ranking individuals within only their own race) to eliminate any average differences in test scores between the races despite differences in skills. Race-norming was attractive to many employers because it lowers validity less (and thus harms productivity less) than low standards for all do. The Civil Rights Act of 1991 banned the practice because it was overtly race conscious (Gottfredson, 1994; Sackett & Wilks, 1994).

Enhance Standards

The second method is to combine a good cognitive test with less cognitive ones that measure job-relevant qualities that cognitive tests do not; for example, noncognitive tests (of personality, interests, etc.) or biographical data blanks (which often contain both cognitive and noncognitive elements). Such supplementation is recognized as the best way to reduce impact because it often raises validity at the same time (Pulakos & Schmitt, 1996). Although cognitive tests best predict the can-do component of job performance (what workers are able to do with sufficient effort), noncognitive tests best predict the will-do component of performance (what they are motivated to do).

The increase in validity gained by using both in combination may or may not

be large, depending on how job related and independent of each other the particular cognitive and noncognitive tests are. Disparate impact falls overall when cognitive tests are supplemented with less cognitive ones because all races score about equally well on noncognitive tests, thus moderating the groups' differences on cognitive tests. However, disparate impact generally does not fall enough to immunize the employer against a legal challenge (Schmitt, Rogers, Chan, Sheppard, & Jennings, in press).

Degrade Standards

The third way of lowering the disparate impact of cognitive tests is to reduce their validity or job relatedness. Tests are not simply either valid or not valid. They vary in the degree to which they predict performance in different occupations. The same principle applies to job performance. Job performance is not just acceptable or not acceptable but ranges on a continuum from abysmal to extraordinary. Successively more valid selection procedures result in successively better performing workforces. Lowering the validity of a hiring procedure thus lowers hiring standards. More valid tests are also fairer to candidates of all races because they more accurately pick the best performers, the most qualified individuals regardless of race.

There are at least three ways of degrading cognitive standards.

Avoid good cognitive tests altogether. This was a common reaction after the *Griggs v. Duke Power Co.* (1971) decision. The test might be replaced by another kind of selection device (say, biographical data inventories). Validity is usually sacrificed in the process, and the drop in workforce performance can be quite marked (Schmidt, Hunter, Outerbridge, & Trattner, 1986).

Use a good cognitive test but in an inefficient way. There are many variants of this strategy. One is to set a low cutoff or pass–fail score, above which all scores are considered equal. This throws away most of the useful information obtained by the test and hence destroys most of its validity. The lower the cutoff, the less useful the test is for identifying the most capable job applicants. *Test-score banding* (Cascio, Outtz, Zedeck, & Goldstein, 1991) is a variant of this. It groups scores into three or more "bands" within which all scores are to be treated as equivalent. Disparate impact can be eliminated or reversed (disfavor the higher scoring group) if the bands are large enough and selection from within bands is race conscious. The loss in validity will depend on the width of the bands and the manner in which individuals are selected from within them.

Another variant is to give a good cognitive test little weight when adding together scores in a battery of tests (cf. Sackett & Wilks, 1994, p. 951, on how employers may "bury" a cognitive test). Some validity will be preserved even with the inefficient use of a good cognitive test, but what remains is mostly the illusion of having measured cognitive skills.

Substitute a poorer test of cognitive skills. Some personnel psychologists have argued that the paper-and-pencil format and abstract nature of traditional cognitive tests impose irrelevant demands on test takers that disadvantage minority test takers. They have therefore sought to develop more concrete tests of mental ability that also mimic what is actually done on the job. These are called *high-fidelity* tests. Hence the popularity at various times of replacing traditional cognitive tests with video-administered exams and job-sample tests. The assump-

tion is that test format and abstractness constitute irrelevant test content and that changing them will reduce disparate impact by removing that irrelevant test content.

This assumption is wrong, however. First, paper-and-pencil format cannot be blamed for disparate impact. The cognitive tests with the greatest disparate impact—intelligence tests—vary greatly in format. Paper-and-pencil tests are only one; orally administered ones requiring neither reading nor writing are another. Moreover, some tests with little disparate impact, including the typical personality test, use the paper-and-pencil format.

Second, abstractness is a highly relevant, not irrelevant, aspect of cognitive tasks. It is the amount and complexity of information that tests require people to process mentally (not whether that information comes in written, spoken, or pictorial form) that create their cognitive demands—and their disparate impact. Mental tasks increase in difficulty and complexity, for example, when there are more pieces of information to integrate, they are embedded in distracting information, and the information is more abstract. This is as true of everyday tasks such as filling out forms and understanding directions as it is of more academic or esoteric tasks (e.g., see Gottfredson, 1997b, on the Educational Testing Service's analysis of items on the National Adult Literacy Survey).

Thus, the more concrete or *contextualized,* well defined, and delimited the tasks on a test, the less complex—and easier—the tests will be. To the extent that high fidelity and other innovative tests do this, they constitute veiled ways of removing relevant demands from cognitive tests. Task difficulty can be leveled and job relatedness lowered in yet other ways, for example, by allowing test takers to take test content home to study (with the help of friends and family) before the exam. The tests may superficially look like good cognitive ability tests, but they are poor substitutes.

It is no surprise, then, that high fidelity is not necessary for job relatedness (Motowidlo, Dunnette, & Carter, 1990) and that nontraditional tests of cognitive ability can reduce validity at the same time they reduce impact (e.g., Pulakos & Schmitt, 1996).

Cognitive tests or their effective use can thus be degraded in various ways and thereby reduce disparate impact. There are many technical decisions in developing selection examinations, each of which can affect the validity of a test to some extent. When those decisions consistently degrade validity for the purpose of reducing disparate impact, the cumulative pattern might be called the *racial gerrymandering of test content.*

Limiting Test Validity in Nassau County

The first and most obvious sign that the Nassau test had been racially gerrymandered was that it excluded precisely what both the literature and its own job analysis indicated it must include—good measurement of cognitive skills. At the same time, the project's technical report (HRStrategies, 1995), curiously, excluded the information necessary to confirm the quality of the test. However, a close reading of the project's account of its technical decisions illuminates how the project had been pressed toward a political purpose.

A Cognitively Empty Test for a Complex Job

The report begins by noting why it is especially important to have a good system for selecting police officers: It "is critical to the safety of the public and reduction of turnover important to proper management of public funds" (HRStrategies, 1995, p. 6). The report's summary of the job, based on the project's extensive job analysis, also makes clear why police work is complex (HRStrategies, 1995):

> [P]atrol officers have primary responsibility for detecting and preventing criminal activity . . . and for enforcement of vehicle and traffic laws. . . . Patrol officers also are charged with responsibility for rendering medical assistance to ill or injured citizens . . . [including] severely injured, mentally ill, intoxicated, violent or suicidal individuals. . . . [They] must pursue ['and take into custody'] individuals suspected of criminal activity . . . [and] have knowledge of the laws and regulations governing powers of arrest and the use of force so as to avoid endangering the public, or infringing upon individuals' rights. . . . Patrol officers . . . must carry out a variety of responsibilities to manage the [crime] scene . . . includ[ing] the identification and protection of physical evidence, identification and initial questioning of witnesses or victims . . . [and] often communicate information they obtain . . . to detectives . . . and others. . . . [They] are regularly assigned to deal with a wide variety of complex emergency situations requiring specialized knowledge and training. . . . In some cases, an immediate, decisive action . . . may be required to protect life or property, or to thwart criminal activity. . . . Patrol officers . . . document extensively their observations and actions . . . and provide statements and court testimony in criminal matters. (pp. 14–15)

Expert police officers from Nassau County then identified 156 "skills, aptitudes, and personal characteristics" that are required for performing well the most important duties in police work. The project ascertained that 106 of them were "critical," 59 of which were "strongly linked" to specific sets of job tasks. Those skills fall into the 18 clusters listed in Table 1. The first 9 clusters are clearly cognitive in nature, the second 9 less so.

The job analysis showed that a variety of skills is critical in police work. As might be expected, however, the Reasoning, Judgment, and Inferential Thinking category turned out to be especially important. Of the 18 categories, it contained the greatest number of both critical skills (17; HRStrategies, 1995, p. 61) and strongly linked ones (13; see Table 1). In addition, unlike all but one other skills category, this one contained skills critical to all duty areas or "task clusters" (HRStrategies, 1995, p. 65–68). As the report describes (HRStrategies, 1995, Suppl. Appendix 4), virtually all large police departments test applicants for judgment/decision-making skills.

The project put together a 25-test experimental battery to measure the 18 types of skills (see Table 1). Not surprisingly, all 3 of the project's centerpiece video-based situation tests, 1 of its 2 paper-and-pencil cognitive tests, and 2 of the 20 personality–temperament measures in the experimental battery were intended to measure reasoning and judgment.

Nonetheless, as shown in Table 1, only one of those six tests remained in the final implementation battery—the personality scale Openness to Experience.

Table 1
Tests Selected to Measure Clusters of Critical Skills

Skill, ability, and personal characteristic cluster	No. of critical skills[a]	Measure in experimental battery
Reading comprehension	1	The list of specific tests used to measure each skills cluster has been omitted here because the publisher declined to give permission to reprint the table as adapted. The omitted information is published in Exhibit 31 from the 1995 project technical report *Nassau County, New York: Design, Validation and Implementation of the 1994 Police Officer Entrance Examination,* by HRStrategies, Detroit, MI. Copyright 1994 by HRStrategies, Inc. Readers can find the exhibit on pages 107–110 of that report.
Reasoning, judgment, and inferential thinking	13	
Listening	1	
Apprehending and restraining suspects	0	
Written communication	7	
Memory and recall	4	
Applying medical procedures	1	
Observation	5	
Oral communication	5	
Cooperation and team work	4	
Flexibility	2	
Creating a professional impression and conscientiousness	7	
Person perception	3	
Vigilance	1	
Willingness to use deadly force	0	
Technical communication	2	
Tools of the trade	1	
Dealing with aided (persons needing aid)	2	

[a]The number of critical skills that were strongly linked to specific sets of task requirements.

Moreover, that scale does not measure the capacity for reasoning and judgment in any way, even according to the project's own definition of the trait ("job involvement, commitment, work ethic, and extent to which work is . . . an important part of the individual's life. . . . [it] includes willingness to work . . . and learn"; HRStrategies, 1995, Appendix S). In short, the project did not measure cognitive ability at all, unless one counts as an adequate cognitive test the ability to read at the level of the bottom 1% of police officers in the research sample. In April 1996, David Jones, president of the consulting firm (HRStrategies) that headed development of the test, concluded a workshop for personnel psychologists (Aon Consulting, 1996a) by stressing that

> the touchstone [of validity] is always back to the job analysis [showing the skills required]. What's in the battery ought to make sense in terms of job coverage, not just the statistics [correlations with on-the-job performance] that come out of the . . . study.

By Jones's own standard, the Nassau test does not measure the skills the job of police officer requires. Nassau County will now be selecting its officers on the

basis of some personality traits with virtually no attention to their mental competence.[1]

Report's Silence on Satisfying the Law

The project had been run by a high-powered group of 10 experts, 5 of them hired by the Department of Justice, who were intimately familiar with both the technical and legal aspects of employee selection. The two leaders of the project's Technical Design Advisory Committee (TDAC) had been appointed by the 1990 consent decree: one to represent the county (David Jones, of HRStrategies) and one to represent the Justice Department (Irwin Goldstein of the University of Maryland, College Park). The former had evaluated or created the county's two previous exams, and the latter is a long-time consultant to the Justice Department on such matters, including earlier litigation in Nassau County.

TDAC's July 1995 technical report (HRStrategies, 1995) is as notable for what it omits and obscures as for what it includes and emphasizes. All such test validation reports should include sufficient information to allow an independent review. The first four pages of the technical report repeatedly stress that it was written to allow a "detailed technical review of the project" (HRStrategies, 1995, p. 2) and even be "understandable to readers not thoroughly familiar with the technology" (HRStrategies, 1995, p. 3). Hundreds of pages and appendixes accompany the 200-page report to facilitate technical review.

However, as shown in Table 2, the report (HRStrategies, 1995) omits most of the crucial information that is required by federal guidelines and recommended by the field of psychology's two sets of professional employment testing standards. TDAC members were fully aware of those standards, many having helped to write them. For example, the report fails to state how well the tests correlated with each other in either the applicant or research groups or, incredibly, even with job performance itself in the research sample of incumbent police officers. It also fails to report how heavily TDAC weighted each test when ranking job applicants. As C. J. Russell (1996) noted, there is "a clear selective presentation of information." The lack of essential information makes it impossible to verify how well scores on the test battery correlated with job performance and thus how job related or valid the exam is.

---

[1]Criterion-related validation studies with police work have produced anomalously low validities for cognitive tests, even when corrected for restriction in range: about .25 versus .51 for comparable jobs (Hirsch et al., 1986; Schmidt, 1997). The occupation is clearly moderately complex, and cognitive ability predicts job performance moderately well at this level of work complexity (e.g., Gottfredson, 1997b; Hunter & Schmidt, 1996; Schmidt & Hunter, in press). It also predicts police academy training performance very well—above .7 (Hirsch et al., 1986). The failure of cognitive tests to correlate more substantially with ratings of police performance on the job may be due largely to problems with the performance ratings. Supervisors have little opportunity to observe police officers performing their duties, meaning that their performance ratings probably are not very accurate.

Low validities of cognitive tests for predicting rated police job performance, therefore, are not a basis for excluding or minimizing their use in police selection. As the *Principles for the Validation and Use of Personnel Selection Procedures* (Society for Industrial and Organizational Psychology [SIOP], 1987, p. 17) state, "The results of an individual validity study should be interpreted in light of the relevant research literature." As already noted, the relevant literature shows that cognitive ability is important for all jobs that, like police work, are at least moderately complex.

Table 2
Major Test Development and Documentation Standards Not Met by the HRStrategies (1995) Report for Nassau County Exam

Uniform Guidelines on Employee Selection Procedures (EEOC et al., 1978)	
15.B.2	Description of existing selection procedures.[a]
15.B.8	Means and standard deviations.[b]
	Intercorrelations among predictors and with criteria.[c]
	Unadjusted correlation coefficients.[d]
	Basis for categorization of continuous data.[e]
15.B.10	Weights for different parts of selection procedure.[f]
Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1985)	
Primary	
1.11	For criterion-related studies, provide basic statistics including measures of central tendency and variability, relationships, and a description of any marked nonnormality of distributions.
1.17	When statistical adjustments made, report both the unadjusted and adjusted results.
6.2	Revalidate test when conditions of test administration changed.
10.9	Give clear technical basis for any cut score.
Secondary	
3.12	Provide evidence from research to justify novel item or test formats.
3.15	Provide evidence on susceptibility of personality measures to faking.
Principles for the Validation and Use of Personnel Selection Procedures (SIOP, 1987)	
Procedures in criterion-related study	
4c	Test administration procedures in validation research must be consistent with those utilized in practice (p. 14).
5d	Use the appropriate formula when adjusting validities with a shrinkage formula (p. 17).
5e	Criterion-related studies should be evaluated against background of relevant research literature (p. 17).
Research reports	
2	Deficiencies in previous selection procedures (p. 29).
9	Summary statistics including means, standard deviations, intercorrelations of all variables measured, with unadjusted results reported if statistical adjustments made (pp. 29–30).
Summary	Provide enough detail in technical report to allow others to evaluate and replicate the study (p. 31).
Use of research results	
12	Take particular care to prevent advantages (such as coaching) that were not present during validation effort. If present, evaluate their effect on validity (p. 34).

Note. EEOC = Equal Employment Opportunity Commission; AERA = American Educational Research Association; APA = American Psychological Association; NCME = National Council on Measurement in Education; SIOP = Society for Industrial and Organizational Psychology.
[a]There are no comparisons of the new procedure with the old procedure. The HRStrategies (1995) report refers readers to the 1988 report that is not attached. [b]These are not reported for the 16 tests winnowed out of experimental battery, for the trial batteries tested or used, or by race for any test. [c]These are not reported for either applicants or incumbents. [d]These are not reported for any of the 25 tests. [e]No basis is given for first percentile reading cutoff. [f]Regression weights are not reported.

Compliance with disparate impact law could have been accomplished with an exam that had either (a) equal validity but less disparate impact than the earlier one or (b) higher validity, whatever its impact. The project clearly set its sights on

satisfying the consent decree by lowering impact rather than raising validity (HRStrategies, 1995, p. 11):

> While the degree of adverse impact for the 1987 examination was less than that experienced with earlier examinations for the position, further *reduction* [italics added] in adverse impact, while *maintaining* [italics added] examination validity, was seen as a key objective of the current project.

However, the project never actually demonstrated that it met this standard either. The report (HRStrategies, 1995) fails to say what either the validity or disparate impact of the 1987 test was and so never demonstrates—or even states—that the 1994 test actually "maintained validity" compared with earlier tests. As seen in Table 2, the federal government's *Uniform Guidelines* (EEOC, 1978, Section 15.B.2) require that "existing procedures" be described, but the report does not do so. Instead, it refers the reader to (but does not attach) the April 1988 report on the previous 1987 exam (a report that the test developers in April 1997 publicly refused to make available to their scientific peers). The project had even included one of the subtests from the 1987 exam (Map Reading) in its experimental battery, specifically to serve as "a benchmark" (HRStrategies, 1995, p. 91) against which to compare the new test and applicant group. Yet, the report never makes any such comparisons. The most the report actually claims is that the validity of the new battery is "statistically significant" (HRStrategies, 1995, p. 135), not that it is equal or superior to earlier tests.

Project Skewed Test Content Away From Good Measurement of Cognitive Skills

TDAC's decisions concerning which tests to include and its justifications for them all worked against cognitive tests and in favor of noncognitive ones. The HRStrategies (1995) report pointedly ignores the large literature on the proven validity of cognitive tests. At the same time, by emphasizing unlikely or disproved threats to their validity and fairness (e.g., paper-and-pencil format), it implies that their use is questionable.

In contrast, a whole appendix is devoted to supporting the validity of personality questionnaires, but no mention at all is made of well-known threats to their validity (e.g., "faking good"). Qualities that many cognitive and noncognitive tests share (which is not pointed out in the report)—such as a paper-and-pencil format—were cited as problematic only in discussing the former. Although cognitive tests of proven general value (traditional ones) were portrayed as narrow and outmoded, the project's unproven substitutes for them were repeatedly extolled as innovative.

No traditional cognitive test was included in the battery, even on a trial basis, except possibly the Map Reading test from the 1987 exam, which soon disappeared from view without comment. One critic complained that "the biggest and most glaring conceptual problem [with the study] is the complete failure to draw on the cumulative scientific literature in any way" (Schmidt, 1996b). Another critic was less charitable: "It seems clear that the authors *did* use prior cumulative knowledge [but] in deciding to minimize the presence of cognitive ability in the predictor domain" (C. J. Russell, 1996).

The HRStrategies (1995) report listed TDAC's four considerations that guided

its decisions about what to include in the experimental battery (pp. 85–86): personality tests, video-administered tests, alternative formats for cognitive tests, and maximum prior exposure to test content and format. All were adopted "in the interests of minimizing adverse impact" (HRStrategies, 1995, p. 86), as Jones has elsewhere suggested that others might do (Aon Consulting, 1996a). By augmenting breadth of coverage, the first could be expected to increase the validity but lower the impact of a test battery containing cognitive tests, but the last three can usually be expected to lower both validity and impact by degrading the validity of the cognitive portion of the exam.

1. *Personality questionnaires.* The project included 20 scales owned by several of the TDAC members (see Table 3): 14 from the *Life Experiences and Preferences Inventory* (LEAP; copyrighted by Personnel Decisions Research Institute) and 6 from the *Work Readiness and Adjustment Profile* (WRAP; copyrighted by Performance Management Associates). The major unresolved question about personality and other noncognitive tests is whether their validity is damaged by job applicants being more motivated to lie or fake good to raise their scores than are the research participants on whom validity is estimated (e.g., Christiansen, Goffin, Johnston, & Rothstein, 1994; Hough, Eaton, Dunnette, Kamp, & McCloy, 1990; Lautenschlager, 1994; Ones, Viswesvaran, & Schmidt, 1993). The report does not mention the faking good issue despite noting a trend in its data that is sometimes thought to signal applicant faking (see Table 3): Applicants got higher scores than police officers on the personality tests (on which lying or faking can raise one's scores) but lower scores, as is usual, on the reading comprehension test (on which lying is useless). Some recent research suggests that faking may not typically be a problem (Ones, Viswesvaran, & Reiss, 1996). However, that optimistic generalization may not apply to Nassau County where the position of police officer is widely coveted for its high pay ($80,000–$100,000 not being uncommon).

2. *Video-based exams.* The project developed three. A Situational-Judgment exercise presented a series of vignettes that portrayed situations in which critical skills are required. Applicants rated how effectively the actor had dealt with the situations enacted. A Learning-and-Applying-Information exercise consisted of a series of video "lessons" about work behavior, which were followed by applicants rating the correctness of an actor's application of that knowledge in pertinent situations. A Remembering-and-Using-Information exercise required applicants to assess whether the behavior of the actor conformed to a fictitious company policy they had been asked to memorize in the month before the exam. None of the three required any reading or writing during the test.

The HRStrategies (1995, p. 85) report described the video exams as having "promise in evaluating applicants' perceptions of complex situations and their approach to dealing with interpersonal activities" in a way that conveys those situations more effectively than a written format but with less disparate impact. No evidence was cited to support this claim. As noted earlier, higher fidelity per se cannot be assumed to improve the valid measurement of cognitive skills.

3. *Alternative formats for measuring cognitive ability.* Among the "promising innovations" the HRStrategies (1995) report suggested for reducing disparate impact without affecting the validity of cognitive tests were "written questions with multiple 'correct' answers or reaction-type responses such as 'agree–

Table 3
The Summary Data Reported for the 25 Tests in the Experimental Battery

Test	Tenure[a]	White–Black difference[b]	Applicant–incumbent difference[c]	Ratio of variances[d]
Situational judgment[e]	−.07	.41	.35	1.05
Remembering and using information	.00			
Learning and applying information	−.03			
Understanding written material	**.12****	**.57**	**−.43**	**1.88**
Reading and using maps	.14**			
LEAP				
Achievement motivation	**−.05**	**.05**	**.56**	**1.01**
Responsibility	**−.16****	**.04**	**.08**	**1.09**
Nondelinquency	**−.21****	**.12**	**.09**	**1.22**
Emotional control	−.27**			
Influence	**.00**	**.09**	**.27**	**1.15**
Sociability	−.09*			
Cooperation	−.23**			
Interpersonal perception	−.02**[f]			
Adaptability	**−.24****	**.11**	**.09**	**1.31**
Tolerance	−.17**			
Fate control	−.10*			
Attention to detail	**.13****	**.07**	**.04**	**.99**
Practical intelligence	−.04			
Authoritarianism (negative)	−.02			
WRAP				
Self-esteem	.09*			
Emotional stability	**−.10***	**−.02**	**.21**	**1.30**
Agreeableness	−.07			
Conscientiousness	.16**			
Openness to experience	**−.09***	**.11**	**.34**	**1.14**
Overall work adaptation	−.08			

Note. Only the tests for the boldfaced numbers were retained in the implementation version of the test battery. The HRStrategies (1995) report provides the last three columns of data only for the 10 tests tried out for the implementation battery. LEAP = Life Experiences and Preferences Inventory; WRAP = Work Readiness and Adjustment Profile.

[a]Correlations of test scores with tenure (HRStrategies, 1995, p. 175). [b]White average minus Black average, in standard deviation units (HRStrategies, 1995, p. 184). The difference is usually about one standard deviation unit for cognitive tests. [c]Applicant average minus incumbent average, in standard deviation units (HRStrategies, 1995, p. 185). [d]Ratio of applicant variance to incumbent variance (HRStrategies, 1995, p. 185). [e]This test was tried out for but not included in the implementation battery. [f]−.02 is not statistically significant, so the HRStrategies (1995) report must be in error at this point.

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

disagree' " and "relaxation of test time limits" (p. 86). All of the video exercises were intended to measure cognitive skills, and two (Remembering and Using Information and Learning and Applying Information) used the agree–disagree format. Ten of the 18 items on the paper-and-pencil cognitive test, Understanding Written Material (discussed below), used the several-correct-answers format. Once again, the project opted for the unproven over the proven in measuring cognitive skills for the purpose of reducing impact.

4. *Maximum exposure of applicants to exam content, format, and requirements in advance of exam.* This was intended to minimize the "test wiseness" that higher scoring groups are often presumed to possess and to benefit from on cognitive tests. Acquainting test takers with test format and requirements is, in fact, good practice because it helps standardize the conditions for valid assessment and minimizes the influence of irrelevant differences among test takers.

Making test content available to applicants beforehand does the opposite. It creates nonstandard conditions that contaminate accurate assessment. Some people will study more or get more assistance from family and friends. It also makes the test much easier by allowing ample time and help for comprehending the materials. The project did this for two exams when it gave applicants the contents up to 30 days before the exam (HRStrategies, 1995, p. 98). One was the video-based Learning-and-Applying-Information test, which required applicants to memorize a fictitious company policy. The second was the paper-and-pencil Understanding-Written-Material test that the project developed to measure reading comprehension. That exam asked applicants questions about reproduced passages of text that they had available for study up to 1 month before the exam.

Moreover, the validation sample of police officers, who were all working full time and not likely to study much, had the materials for only 1 week. Thus, test-taking conditions were not standard among the applicants and they differed between the applicant and research groups too, which clearly violates both good practice and professional testing standards (e.g., Standards 4c and 12 of the *Principles for the Validation and Use of Personnel Selection Procedures* [*SIOP Principles*]; SIOP, 1987; see Table 2).

Interestingly, when two TDAC members had been retained to evaluate the 1983 Nassau exam, they had recommended throwing out the scores for almost half of the questions on the exam (its "book" questions) precisely because applicants had been given exam material to study 2 weeks before the test: "A Pre-Examination Study Booklet with unknown influence on individual test performance was used, thus compromising standardization of a significant portion of this test" (Jones & Prien, 1986, p. II.3).

In summary, the project used two of the three procedures outlined earlier that reduce disparate impact by degrading the valid measurement of cognitive skills: omitting cognitive tests with proven validity and substituting nontraditional ones of uncertain validity. As is shown, the project would later use the third strategy too (inefficient use of cognitive scores) by regrading the reading comprehension test pass–fail with the passing score set at the lowest possible level. As C. J. Russell (1996) noted, the "major impression . . . [is that] all decisions in the Nassau study were driven by impact adjustments."

Project Tilted Validity Calculations Against Cognitive Tests and in Favor of Noncognitive Ones

The project next evaluated how well the scores on the 25 experimental tests related to the job performance ratings of 508 Nassau County police officers. The objective was to identify the most useful tests for inclusion in a final implementation test battery for ranking applicants. The HRStrategies (1995) report states (but never shows) that all tests with significant validity were retained, for a total of 10:

8 of the personality scales, the video-based Situational Judgment, and the paper-and-pencil Understanding-Written-Material test (see Table 3).

The project made some odd and unexplained decisions in this winnowing process. First, TDAC winnowed the 25 tests in a peculiar manner (HRStrategies, 1995, pp. 130–133), too obscure to explain fully here. Briefly, it involved retaining only those tests that TDAC had predicted would be related in highly particular ways to different dimensions of job performance. While ostensibly intended to minimize a technical problem ("capitalizing on chance"), this procedure would have allowed TDAC prejudices and misconceptions about which outcomes the cognitive tests would predict to influence its decisions about which tests to retain. The report provides data on neither the job relatedness nor the disparate impact of the 15 tests eliminated at this point, violating all three sets of test standards in the process (see Table 2).

This curious procedure and the missing data are especially troubling in view of a second odd decision, which the report itself characterized as "unique": to administer the 25-test experimental exam to the 25,000 applicants before validating it (HRStrategies, 1995, p. 7). This decision, which reverses the usual sequence of first establishing validity among incumbents and then administering the (valid) test to applicants, "would afford noteworthy research advantages with regard to exploring and creating a 'potentially less adverse alternative' selection device" (HRStrategies, 1995, p. 119). Its advantage would be that "the research team could view the operation of creative examination formats within a true applicant group, prior to eliminating components which might appear to work *less effectively* [italics added] if viewed solely from the perspective of a concurrent, criterion-related [job performance-related] validation strategy" (HRStrategies, 1995, p. 7).

Translated, this means that TDAC wanted first to see the disparate impact of different tests in its experimental battery so that it did not inadvertently commit itself to using tests with substantial disparate impact even if they had the highest validity or, conversely, to omitting less valid tests if they had favorable racial results. The HRStrategies (1995) report repeated this reason on the next page in implicitly justifying why applicants had been given tests (about four hours' worth) that did not actually count toward their scores. TDAC has since claimed (Dunnette et al., 1997) that the reversal in procedure was meant to protect test security, but the report itself gives no hint of any such concern, emphasizing instead TDAC's goal of reducing disparate impact.

Third, the correlations used in showing the job relatedness of different tests and test combinations were calculated in a way that could be expected to suppress the apparent value of cognitive tests relative to noncognitive ones. As C. J. Russell (1996) has noted, "We see the authors bending over backwards to eliminate cognitive test remnants from the predictor domain."

The project did not report the usual unadjusted (zero-order) correlations required by all three sets of test standards (see Table 2) but instead reported the twice-adjusted ones that the project called *simple validities*.[2] By omitting the

[2]The project had statistically partialed tenure (length of experience on the police force) out of both the predictors (test scores) and criteria (performance ratings). While not viewed favorably by some test developers, partialing tenure out of the criterion performance ratings is not unusual as a means of controlling for differences in job experience. More experienced workers tend to perform

required unadjusted correlations, TDAC had made it impossible for others to verify the predicted differential tilting of results. However, when pressed, TDAC recently provided some of the missing results (Dunnette et al., 1997), and they confirmed the prediction of tilted results. TDAC's adjustments made the noncognitive tests appear 35% more valid than the cognitive ones when, in fact, the average unadjusted validities for both test types were equal.[3]

Those just-revealed unadjusted correlations also point to the foolhardiness of administering a battery of unproven innovative tests to 25,000 applicants before assessing their worth: Their validities were shockingly low, for an average (absolute value) of only .05 (on a scale from 0 to 1.00). Only 3 of the 25 tests had validities reaching .10. Worthless or not, the project had already committed the county and its applicants to the test.

---

better because they learn on the job, and this suppresses the apparent validity of the useful traits (like cognitive ability) that they bring with them into the job but that do not change with experience. However, the project partialed tenure out of the predictors as well, but there is no theoretical reason to do so and the report gives none. The problem is this.

As shown in Table 3, tenure is positively correlated with the more cognitive tests and negatively with all but one personality scale. The HRStrategies (1995) report itself suggested that the more experienced officers had been selected under different standards (p. 131), which helps explain why they did better on the cognitive tests than less experienced officers. (Nassau County's hiring standards seem to have fallen in recent years because consent decrees degraded both its 1983 and 1987 exams.) TDAC's report seems to argue that such changes in standards require controlling for tenure, when, in fact, they mean that tenure-related differences in performance are not related to experience and therefore should not be controlled. Partialing tenure out of the predictors thus amounted to partialing some of the valid variance out of the cognitive tests. This would depress their apparent correlation with job performance. On the other hand, partialing tenure out of the predictors would raise the apparent value of the noncognitive tests because they were negatively correlated with tenure (see Table 3).

It might also be noted that partialing tenure out of the criterion may not have been entirely appropriate in the current situation. As noted above, more experienced officers tended to score higher on the cognitive tests, but this is unusual. Because ability was correlated with tenure among Nassau police officers, controlling for tenure in the criterion will necessarily at the same time partial out some of the valid covariance between the cognitive tests and the criterion, even though that was not its purpose. That is, some of the correlation of tenure with job performance is spurious because of tenure's correlation with a known cause of superior job performance—cognitive ability.

This problem can be better visualized by noting that today's tenure will correlate with yesterday's training performance in Nassau County (which obviously cannot be a causal relation) simply because earlier trainees were brighter on average than more recent ones. (Mental ability is a good measure of trainability.) Partialing tenure out of training grades would obviously be inappropriate because their relation with tenure is entirely spurious. Although not entirely spurious, the correlation between tenure and incumbents' job performance is partly so in Nassau County.

[3]TDAC (Dunnette et al., 1997) has argued that the tenure adjustment made no difference, but it invokes only irrelevant statistics to support its claim (Gottfredson, 1997a). The pertinent statistics show that the adjustment made considerable difference. Before the scores were adjusted, job-relatedness correlations were the same on the average for the two cognitive tests as for the eight personality tests—.08 (on a scale from 0 to 1.00). Adjusting the job performance ratings (the criterion) for tenure raised correlations for the noncognitive tests (to .095) and lowered them for the cognitive tests (to .075). This made the apparent validity of the personality tests 27% larger than that of the cognitive tests. Controlling for tenure in test scores as well as in criteria ratings increased the gap to 35% by boosting the noncognitive correlations a bit beyond .10.

Because all of the correlations were so low, another advantage of the double adjustment was simply to raise the apparent validity of most of the tests.

Project Kept Little More Than the Illusion of Testing for Cognitive Ability

The project next considered which of the remaining 10 tests it would use, and how, in the implementation battery. It tried out five "basic" prediction models with different combinations of the 10 tests, 4 of which included at least 1 of the 2 putatively cognitive tests (the video-based Situational Judgment and the paper-and-pencil Understanding Written Material). Having first degraded the cognitive parts of the experimental battery and then understated their job relatedness, the project not surprisingly found that the five models yielded "nearly identical" validities (HRStrategies, 1995, p. 135) whether or not they contained a cognitive test (Table 4 shows the results for several). The project was now free to rest its decision entirely on the alternative batteries' disparate impact. The battery with the least impact was the noncognitive model consisting solely of personality scales.

However, TDAC balked at recommending it—and rightly so—despite it being the only one to meet for Blacks the federal government's four-fifths rule. (The federal government's rule of thumb is that disparate impact is present and can trigger litigation when the proportion of a minority group's applicants who are selected is less than four fifths the proportion of Whites selected.) The HRStrategies (1995) report states that "TDAC was concerned that implementation of this battery, containing no formal measure of reading comprehension or other cognitive skills, could potentially admit applicants to Police Academy training who would

Table 4
Estimates of Validity of Alternative Prediction Models

Model	Observed	Shrunken	Corrected for: Criterion unreliability	Corrected for: Range restriction	Impact ratio[a]
Reported in HRStrategies (1995) report					
All 25 predictors	.30	.20	.25	—	—
Full model (eight noncognitive scales, written material, and situational judgment)	.24	.20	.25	.31	.62
Noncognitive (eight noncognitive scales)	.22	.20	.25	.29	.82
Refined model (eight noncognitive scales and first percentile reading on written material test)	.23	.20	.25	.35	.77
Re-estimated by Schmidt (1996b)					
Refined model					
Minimum[b]		.05		.08[c]	
Maximum[d]		.14		.20	
Best estimate[e]		.10		.14	

Note. Dashes indicate that data were not reported.
[a]Disparate impact ratio (percentage of Blacks passing divided by percentage of Whites passing). [b]Based on shrinking the average of the observed validities for all six models in the report, .228. [c]This column corrected for both unreliability and restriction in range. [d]Based on shrinking the observed validity of the 25-variable regression model, .25. [e]The average of the minimum and maximum estimates.

fail in the training program" (p. 139; see also Goldstein's court testimony, *United States v. Nassau County*, 1995b, p. 65). Suddenly there is a glimpse of TDAC's knowledge of the literature concerning cognitive ability showing that general mental ability is the major determinant of trainability (e.g., Gottfredson, 1997b; Hirsch, Northrop, & Schmidt, 1986; Hunter & Hunter, 1984; Rafilson & Sison, 1996) but that personality plays a smaller role (e.g., Ones & Viswesvaran, 1996; Schmidt & Hunter, in press). TDAC's solution was to restore the reading test—but rescored with the passing score set at the first percentile of incumbent officers. This was the project's hybrid or refined model.

TDAC gives no rationale for dichotomizing the reading scores, as is required by the test standards (e.g., 15.B.8 of the *Uniform Guidelines* [EEOC et al., 1978] and 6.9 and 10.9 of the *Standards for Educational and Psychological Testing* [AERA, APA, & NCME, 1987]). Nor does it attempt to give a technical rationale for such a dramatically low cutoff, which no doubt minimized the reading test's disparate impact. The HRStrategies (1995) report says only that TDAC "assume[d] that applicants scoring at or below this level [the incumbents' first percentile] might represent potential 'selection errors' " (p. 139).[4] In short, TDAC pulled back only slightly from completely eliminating all cognitive demands from the exam.

Three Mistakes Inflated the Apparent Validity of the Cognitively Denuded Implementation Battery

Intentionally or not, TDAC had systematically denuded its final test battery of most cognitive content, which could be expected to damage the exam's validity. That damage is not apparent in the technical report (HRStrategies, 1995), however, because TDAC made three statistical errors that inflated the battery's apparent merit relatedness by over 100%. All three errors occurred in correcting the test battery's correlation with job performance for two of three statistical artifacts that are known to distort such correlations in predictable ways. The first artifact (capitalization on chance) artificially inflates the apparent job relatedness of a battery of tests (its overall correlation with job performance ratings); the second and third artifacts (criterion unreliability and restriction in range on the predictors) artificially depress apparent job relatedness. Correcting for the three artifacts results in a more accurate estimate of how useful a test battery will be when it is actually used to hire new workers (what is technically called its *true validity*).

To correct for the first artifact, the project applied a "shrinkage" formula to the correlation calculated for the test battery in the research sample. This is the less

[4]Justice's Gadzichowski has dismissed criticism of the low reading minimum as "uninformed and unfounded" (July 25, 1996 letter from John M. Gadzichowski to Frank Erwin). Justice, like TDAC (Dunnette et al., 1997), has defended the minimum by arguing that the five officers who scored lowest on the reading test must be competent because they all had at least 2 years of college credit. If police department anecdotes are correct, however, accumulating 2 years of college credits does not assure competence in filling out even the simplest incident forms. Nor would one expect it to in view of the fact that in the United States virtually anyone can take courses at some sort of college. The U.S. Department of Education's 1993 National Adult Literacy Survey shows, in fact, that fully 4 percent of college graduates comprehend written material no better than the bottom one quarter of all adults in the United States (National Adult Literacy Survey Level 1 out of 5; Kirsch et al., 1993, pp. 116–118), which is a literacy level far, far below what police work requires.

preferred but sometimes necessary route when a project includes in its test battery only some of the tests it tried out. Although not necessary in this case, the use of a shrinkage formula allowed TDAC to make two errors that resulted in shrinking its correlation far too little. TDAC's first error was to shrink the wrong, much higher correlation of .30 (from the 25-test battery) instead of .23 (for the 9-test refined battery). In other words, TDAC gave the 9-test battery credit for being as predictive as the 25-test battery, which it clearly was not.

Second, TDAC applied the wrong shrinkage formula, which shrunk that already too-high correlation by too little.[5] This latter error was particularly puzzling because one TDAC member had written an article some years earlier on avoiding the error (Schmitt, Coyle, & Rauschenberger, 1977). The *SIOP Principles* (SIOP, 1987) are explicit, moreover, in requiring the "appropriate shrinkage formula" (Standard 5d in Table 2). (TDAC has since admitted this error; Dunnette et al., 1997.) The same two errors were made for the other five combinations of tests that the project tried out.

Having failed to shrink the correlation for its six alternative batteries far enough downward to correct for the first artifact, the project then adjusted too far upward the correlation for its favored refined battery when correcting for the third artifact.[6] Thus, while TDAC had ballooned the apparent validity of all the alternatives it tested for the final battery, it inflated even further the apparent value of its preferred alternative.

Schmidt (1996b) estimated that the project's first two statistical errors improperly inflated the "true" validities for all six trial batteries by at least 100%. Lacking the data to recalculate them, he derived minimum and maximum estimates (see Table 4). TDAC had estimated the true validity of its recommended

---

[5]Regression models (for calculating the multiple correlation of a set of tests with job performance) always capitalize on chance by delivering the best fit possible to the data in hand, chance factors and all. This means that validities estimated in the research sample are always somewhat inflated. The best solution for deriving a more accurate (smaller) estimate is to apply the regression weights developed in the research sample to an independent cross-validation sample that was not involved in selecting the battery. The Nassau project instead used a shrinkage formula to adjust the observed validities of its alternative prediction models.

According to Schmidt (1996b), however, it used the wrong shrinkage formula (the Wherry correction instead of Cattin's, 1980, Equation 8), which provides too large an estimate when the validity to be shrunk is from a regression model excluding some of the original variables in the study, as was the case here. TDAC then applied this mistaken formula to the wrong validity—the multiple correlation for the regression equation including all 25 variables (.30) that, as can be seen in Table 4 (column 1), is considerably larger than the validity observed for any of the models actually being tested (.22–.24). It then assigned that single, too-large shrunken validity (.20) to all of the models.

[6]Observed validities are often corrected for criterion unreliability (third column in Table 4) and restriction in range on the predictors (fourth column). The project made these two corrections, as is appropriate in typical circumstances. However, the estimated true validity for its preferred refined model (.35) is clearly mistaken. The full model contains all nine tests that are in the refined model (plus one more), and its observed validity (.24) is essentially the same as for the latter (.23). It therefore makes no sense that the correction for restriction in range would boost the latter's estimated true validity by almost twice as much—.12 (from .23 to .35) versus .07 (from .24 to .31)—when virtually the same tests are involved. Nor does it make sense that the model with the less efficient (pass–fail) use of the reading test would produce the higher validity (.35 vs. .31) for the very same people. The HRStrategies (1995) report does not describe how it carried out the corrections, but the project probably made an error in correcting for restriction in range for the dichotomized reading scores in the refined model. (Table 3 shows degree of restriction for all the predictors.)

battery to be .35 (on a scale from 0 to 1.0), but Schmidt estimated it to be less than half that—about .14.

Finally, it must be remembered that the foregoing estimates were based on the project's improperly doubly adjusted "simple" correlations, which themselves were probably inflated for the noncognitive tests that dominated the final battery. In fact, one might wonder whether those improper simple correlations, by tilting the correlations against the cognitive tests and in favor of the noncognitive ones, might have created some anomalies in how those prediction models weight the different tests. Those regression weights, however, were not reported as required by the *Uniform Guidelines* (EEOC et al., 1978, 15.B.10).

Incorrect Testimony Misleads Judge

Justice's Gadzichowski (*United States v. Nassau County,* 1995b, p. 23) testified that the new exam not only had less disparate impact than the 1987 test but was also twice as valid. His numbers were .35 for the new test versus .12 (or .165 after "modification") for the earlier one. However, not only was the .35 a grossly inflated estimate, but it was the wrong statistic (and highly favorable) for the comparison at hand. Gadzichowski had compared the erroneously estimated true validity of the 1994 exam (.35) with the necessarily much lower observed validity of the 1987 exam (about .12–.16). Two TDAC members were present during Gadzichowski's testimony but did not correct his improper comparison. Although Gadzichowski did not report the 1987 exam's estimated true validity, it is probably higher than the new exam's because the former's observed validity (.12–.16) is as high as the new test's true validity (.14) when properly estimated (see Table 4).

Gadzichowski also compared the new exam favorably with the 1983 exam. A decade earlier, two TDAC members (Jones & Prien, 1986, p. VIII.9) had reported the observed and true validities of the 1983 exam to be, respectively, .22 and .46 (.21 and .40 if the "book" questions were omitted as they recommended). Schmidt's (1996b) best estimate of the 1994 exam's true validity (.14) indicates that it is far less job related than the 1983 exam (.40 or more).[7] It is also less valid than the typical cognitive test in police work—about .25, which is probably an underestimate (see Hirsch et al., 1986; Schmidt, 1997).

Nevertheless, the Court, operating on what it had been told, approved the new exam for use in Nassau County at the conclusion of the hearing at which Gadzichowski testified.

Implications

The Nassau County police exam may be no more valid for selecting good police officers than flipping a coin. If at all valid, it is considerably less so than at

[7]The Justice Department might argue that the validity of the 1983 exam was actually zero, not the .2 (observed) and .4 (true) that Jones and Prien (1986) had estimated. The reason is that Justice had apparently allowed civil rights lawyers to pick apart the 1983 and 1987 exams so that they could (improperly) challenge their validity. By breaking a reliable test into its necessarily less reliable pieces or by breaking a research sample into many small groups, it is always possible to capitalize on chance factors to seem to show that some aspect of the test is not valid for some segment of the population. Such opportunistic data ransacking in fact enabled civil rights lawyers to convince the District Court that they should be allowed to rescore the 1983 and 1987 tests to reduce disparate impact (*United States v. Nassau County,* 1995b, p. 15).

least one of the county's two earlier tests and less than the ones now used by many other police departments around the country. The Justice Department has thus forced the county, perhaps unlawfully, to lower its standards in the guise of improving merit hiring. Also, TDAC has provided Justice with scientific cover for doing so.

Nassau County

The millions of dollars Nassau County was forced to spend for the new test are only the first of the costs the test will impose on the county. Because the test is less effective than earlier ones in screening for mental competence, Nassau County will either see a rising failure rate in training or else be forced to water down academy training. Job performance will also fall as new classes of recruits make up a bigger segment of the police force and move into supervisory positions. If Washington, DC's experience with lax standards is any guide, complaints of police brutality will rise, lives and equipment will be lost at higher rates, and the credibility of the force will fall (Carlson, 1993).

The county might once have been able to rely on educational credentials to maintain its standards, but it cannot now. Although not mentioned in TDAC's report, the Justice Department forced the county some years ago to abandon its requirement for 2 years of college. Justice's current consent decree with the county allows it to require only 1 year of college credits—and then only if that requirement has no disparate impact.

This twin lowering of cognitive standards comes, moreover, when the Nassau County Police Department has just introduced community policing into its eight precincts. Problem solving or community policing is a new model for policing that is being adopted by progressive departments throughout the country (e.g., Goldstein, 1990; Sparrow, Moore, & Kennedy, 1990). Former Attorney General Edwin Meese, III (1993, p. 1) described how the new policing changes the fundamental nature of police work:

> Instead of reacting to specific situations, limited by rigid guidelines and regulations, the officer becomes a thinking professional, utilizing imagination and creativity to identify and solve problems ... [and] is encouraged to develop cooperative relationships in the community.

By maximizing individual officers' participation in decision making, it creates even higher demands for critical thinking and good judgment. The new test, stripped of most cognitive content, will doom realization of this new vision of policing in Nassau County.

Nassau County loses not only the benefit of the many talented people it might otherwise have been able to hire but also its legitimacy as a fair unit of government. Highly qualified people of all races lose job opportunities that should have been theirs under merit hiring. They learn that talent, hard work, and relevant experience no longer count for much.

U.S. Justice Department

This case study illustrates how Justice's Civil Rights Division is enforcing a political agenda of its own making, usurping for itself the powers arrogated to Congress. By degrading merit hiring, it also works against the administration's

own programs (e.g., Community-Oriented Policing Services Program and Police Corps) for improving the quality of policing nationwide.

Disparate impact may be the trigger for legal action, but it is not the ultimate standard for the lawfulness of a selection procedure. Validity is (EEOC et al., 1978, Questions 51 and 52). Under the law, validity trumps disparate impact. Not so for the Justice Department, however, whose yardstick is clearly disparate impact and for whom validity has been mostly an impediment in pursuing its goal of no impact.

This case also raises a new question about civil rights law. Is it illegal to craft the contents of a test to favor some races or disfavor others when such procedures artificially cap or lower the test's validity? For example, does it constitute intentional discrimination to exclude good tests from a battery simply because proportionately more Whites than Blacks do well on them? Or to rescore and degrade a test battery, after the fact, solely to increase the number of Blacks who pass it? Section 106 of the Civil Rights Act (1991) forbids the race-conscious adjustment of test scores, so it would seem to follow that race-conscious adjustment of test content to engineer racial outcomes would also be proscribed. In addition, Section 107 of the act states that race cannot be "a motivating factor" in selecting employees.

A related matter that Congress might investigate is whether the Justice Department's involvement in developing and promoting tests compromises its ability to enforce the law impartially and impermissibly interferes with competition in the test marketing business. Is there not a conflict of interest when the Justice Department is asked to litigate a test that it helped develop? Was there not a conflict of interest for Justice's Gadzichowski to dispute the merits of the *Hayden et al. v. Nassau County* (1996b) lawsuit alleging reverse discrimination in the new test?

Despite its claims to the contrary, the Justice Department has been recommending particular tests and test developers over others. Its involvement with Aon Consulting, both in Nassau County and in Aon's recent test validation consortium, gives Aon an enormous advantage over other test developers, whatever the quality of its product. Test developers around the country report that they have begun to lose business because of Justice Department pressure on their clients to use some variant of the Nassau test. That pressure has included the Department of Justice making extraordinary demands on police departments for information and threatening to sue or to refuse to end a consent decree. For many jurisdictions, a Justice Department suggestion is clearly an offer they cannot refuse.[8]

Psychology

Both employment discrimination law and Justice Department enforcement of it are premised on assumptions that contradict scientific knowledge and professional principles in personnel psychology. As some have said, psychometricians are expected to be "psychomagicians"—to measure important job-related skills without disparate impact against the groups that possess fewer of the skills.

[8]The Constitution Subcommittee of the House Judiciary Committee recently became interested in the Department of Justice's involvement in police testing. In a May 20, 1997, oversight hearing on the Civil Rights Division, the Subcommittee heard testimony on improper justice action in two such cases (*Testimony of W. Flick and L. S. Gottfredson*).

Lacking magic, psychologists are tempted to appear to have worked it nonetheless. The Justice Department and many employers expect nothing less. The result may be compromise (reduce disparate impact by reducing validity) or capitulation (eliminate disparate impact regardless of what is required). However, in either case, sacrificing validity for racial reasons constitutes a covert political decision on the part of the psychologist if done without reviewing all options with the employer.

Some psychologists have suggested that validity be lowered somewhat to reduce disparate impact in the name of balancing social goals (Dunnette et al., 1997; Hartigan & Wigdor, 1989; Zedeck, Cascio, Goldstein, & Outtz, 1996). This is a legitimate political position about which personnel psychologists possess relevant information. However, such positions, whether explicit or not, are political and not scientific. They need to be aired in the political arena, not enacted covertly or in the name of science. Only with public airing of the trade-offs involved will unreasonable employment discrimination law and enforcement be revealed for what they are, perhaps relieving some of their corrupting pressure on selection psychologists to perform psychomagic.

Every test developer who manipulates content to reduce disparate impact lends credence to the egalitarian fiction that, but for discrimination, all demographic groups would be hired in equal proportion in all jobs. It does so by appearing to reduce or eliminate disparate impact without race-conscious selection, thus concealing the real dilemmas that bedevil work in this area. The illusion of easy success in substantially eliminating disparate impact makes it more difficult for honest developers to get business and for employers to withstand pressure to eliminate racial disparities at any price. The absence of overt race consciousness also removes any obvious basis for alleging reverse discrimination, as Nassau County plaintiff William Hayden and his colleagues discovered.

The technical report (HRStrategies, 1995) for the 1994 Nassau County police test suggests that TDAC's efforts were bent to the political will of the Justice Department and provided technical camouflage for that exercise of will. Psychologists might ponder under what conditions they should even participate in such "joint" projects in which there is confusion about who the client really is and in which one partner has the power to harass and punish the other with impunity. The ethics of independent psychologists working jointly with the Justice Department (with Justice Department "oversight") become even murkier when the relation with Justice is a long-term, lucrative one spanning a series of not-entirely voluntary clients to whom Justice provides the firm "access" by its much-flexed power to intimidate.

Psychology could do at least two things to help its practitioners avoid becoming compromised in personnel selection work. One is to clarify the ethical considerations that should govern contracts involving both clients and the enforcement agencies to which they are subject. Another is to clarify—publicly—the counterfactual nature of employment discrimination law and the rogue nature of its enforcement by the Justice Department.

References

American Educational Research Association & American Psychological Association. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Aon Consulting. (1996a). *EEO legal and regulatory developments* [Video]. Detroit, MI: Author.

Aon Consulting, (1996b). *HRStrategies entry-level law enforcement selection procedure design and validation project.* Detroit, MI: Author.

Brody, N. (1996). Intelligence and public policy. *Psychology, Public Policy, and Law, 2,* 473–485.

Carlson, T. (1993, November 3). Washington's inept police force. *Wall Street Journal,* p. A23.

Cascio, W. F., Outtz, J., Zedeck, S., & Goldstein, I. L. (1991). Statistical implications of six methods of test score use in personnel selection. *Human Performance, 4,* 233–264.

Cattin, P. (1980). Estimating the predictive power of a regression model. *Journal of Applied Psychology, 65,* 407–414.

Christiansen, N. C., Goffin, R. D., Johnston, N. G., & Rothstein, M. G. (1994). Correcting the 16PF for faking: Effects on criterion-related validity and individual hiring decisions. *Personnel Psychology, 47,* 847–860.

Civil Rights Act of 1991, Pub. L. No. 102–166, §§106–107, 105 Stat. 1071 (1991).

Dunnette, M., Goldstein, I., Hough, L., Jones, D., Outtz, J., Prien, E., Schmitt, N., Siskin, B., & Zedeck, S. (1997). *Responses to criticisms of Nassau County test construction and validation project.* Unpublished manuscript. Available www.ipmaac.org/nassau/

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice. (1978). Uniform guidelines on employee selection procedures. *Federal Register, 43* (166).

Goldstein, H. (1990). *Problem-oriented policing.* Philadelphia: Temple University Press.

Gottfredson, L. S. (1986). Societal consequences of the *g* factor in employment. *Journal of Vocational Psychology, 29,* 379–410.

Gottfredson, L. S. (1994). The science and politics of race-norming. *American Psychologist, 49,* 955–963.

Gottfredson, L. S. (1996a, December 10). New police test will be a disaster [Letter to the editor]. *Wall Street Journal,* p. A23.

Gottfredson, L. S. (1996b, October 24). Racially gerrymandered police tests. *Wall Street Journal,* p. A18.

Gottfredson, L. S. (1996c). *The hollow shell of a test: Comment on the 1995 technical report describing the new Nassau County police entrance examination.* Unpublished manuscript, University of Delaware. Available www.ipmaac.org/nassau/.

Gottfredson, L. S. (1997a). *TDAC's defense of its Nassau County police exam makes my point.* Unpublished manuscript, University of Delaware. Available at www.ipmaac.org/nassau/.

Gottfredson, L. S. (1997b). Why *g* matters: The complexity of everyday life. *Intelligence, 24,* 79–132.

Griggs v. Duke Power Co., 401 U.S. 424 (1971).

Hartigan, J. A., & Wigdor, A. K. (Eds.). (1989). *Fairness in employment testing: Validity generalization, minority issues, and the General Aptitude Test Battery.* Washington, DC: National Academy Press.

Hayden et al. v. Nassau County, N.Y. Trial/I.A.S. Part 13, Index No. 14699/96 [Affirmation in opposition] (1996a, June 6).

Hayden et al. v. Nassau County, N.Y. Trial/I.A.S. Part 13, Index No. 14699/96 [Motion] (1996b, July 1).

Hirsch, H. R., Northrop, L. C., & Schmidt, F. L. (1986). Validity generalization results for law enforcement occupations. *Personnel Psychology, 39,* 399–420.

Hough, L. M., Eaton, N. K., Dunnette, M. D., Kamp, J. D., & McCloy, R. A. (1990). Criterion-related validities of personality constructs and effect of response distortion on those validities [Monograph]. *Journal of Applied Psychology, 75,* 581–595.

HRStrategies. (1995). *Nassau County, New York: Design, validation and implementation*

of the 1994 police officer entrance examination (Project technical report). Detroit, MI: Author.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96,* 72–98.

Hunter, J. E., & Schmidt, F. L. (1996). Intelligence and job performance: Economic and social implications. *Psychology, Public Policy, and Law, 2,* 447–472.

Hunter, J. E., Schmidt, F. L., & Rauschenberger, J. (1984). Methodological, statistical, and ethical issues in the study of bias in psychological tests. In C. R. Reynolds & R. T. Brown (Eds.), *Perspectives on bias in mental testing* (pp. 41–99). New York: Plenum.

Jones, D. P., & Prien, E. P. (1986, February). *Review and criterion-related validation of the Nassau County Police Officer Selection Test (NCPOST).* Detroit, MI: Personnel Designs.

Kirsch, I. S., Jungeblut, A., Jenkins, L., & Kolstad, A. (1993). *Adult literacy in America: A first look at the results of the National Adult Literacy Survey.* Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.

Lautenschlager, G. J. (1994). Accuracy and faking of background data. In G. S. Stokes, M. D. Mumford, & Owens, W. A. (Eds.), *Biodata handbook: Theory, research, and use of biographical information in selection and performance prediction* (pp. 391–419). Palo Alto, CA: Consulting Psychology Press.

Meese, E., III. (1993). Community policing and the police officer. In S. Michaelson (Ed.), *Perspectives on policing* (No. 15, pp. 1–11). Washington, DC: U.S. Department of Justice, National Institute of Justice, & Harvard University, Kennedy School of Government.

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology, 75,* 640–647.

NAACP and New Jersey Conference NAACP v. State of New Jersey, Department of Law and Public Safety, Division of State Police, No. MER-L-002687-96 (N.J. Sup. Ct. June 24, 1996).

Nelson, M., & Shin, P. H. B. (1994, August 1). Testers' bad mark. *Newsday,* pp. A5, A22.

O'Connell, R. J., & O'Connell, R. (1988). Las Vegas officials charge Justice Department with coercion in consent decrees. *Crime Control Digest, 22*(49), 1–5.

Ones, D. S., & Viswesvaran, C. (1996, April). *A general theory of conscientiousness at work: Theoretical underpinnings and empirical findings.* Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, San Diego, CA.

Ones, D. S., Viswesvaran, C., & Reiss, A. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology, 81,* 660–679.

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

Oversight hearing regarding the Civil Rights Division of Department of Justice Committee on the Judiciary: Hearing testimony presented to the subcommittee on the Constitution, 105th Cong., 1st session (May 20, 1997) (testimony of W. Flick and L. S. Gottfredson).

Pulakos, E. D., & Schmitt, N. (1996). An evaluation of two strategies for reducing adverse impact and their effects on criterion-related validity. *Human Performance, 9,* 241–258.

Rafilson, F., & Sison, R. (1996). Seven criterion-related validity studies conducted with the National Police Officer Selection Test. *Psychological Reports, 78,* 163–176.

Russell, C. J. (1996). *The Nassau County police case: Impressions.* Unpublished manuscript, University of Oklahoma. Available at www.ipmaac.org/nassau/.

Russell, T. L., Reynolds, D. H., & Campbell, J. P. (1994). *Building a joint-service*

classification research roadmap: Individual differences measurement (Tech. Rep. No. AL/HR-TP-1994-0009). Brooks Air Force Base, TX: Armstrong Laboratory.

Sackett, P. R., & Wilks, S. L. (1994). Within-group norming and other forms of score adjustment in preemployment testing. *American Psychologist, 49,* 929–954.

Schmidt, F. L. (1988). The problem of group differences in ability test scores in employment selection. *Journal of Vocational Behavior, 33,* 272–292.

Schmidt, F. L. (1996a, December 10). New police test will be a disaster [Letter to the editor]. *Wall Street Journal,* p. A23.

Schmidt, F. L. (1996b). *Some comments on the Nassau County police validity case.* Unpublished manuscript, University of Iowa. Available www.ipmacc.org/nassau/.

Schmidt, F. L. (1997). *Comments on the 1997 SIOP symposium on the Nassau County police test.* Unpublished manuscript, University of Iowa. Available at www.ipmaac.org/nassau/.

Schmidt, F. L., & Hunter, J. E. (in press). Measurable personnel characteristics: Stability, variability, and validity for predicting future job performance and job related learning. In M. Kleinmann & B. Strauss (Eds.), *Instruments for potential assessment and personnel development.* Gottingen, Germany: Hogrefe.

Schmidt, F. L., Hunter, J. E., Outerbridge, A. N., & Trattner, M. H. (1986). The economic impact of job selection methods on size, productivity, and payroll costs of the federal work force: An empirically based demonstration. *Personnel Psychology, 39,* 1–29.

Schmitt, N., Coyle, B. W., & Rauschenberger, J. (1977). A Monte Carlo evaluation of three formula estimates of cross-validated multiple correlation. *Psychological Bulletin, 84,* 751–755.

Schmitt, N., Rogers, W., Chan, D., Sheppard, L., & Jennings, D. (in press). Adverse impact and predictive efficiency using various predictor combinations. *Journal of Applied Psychology.*

Sharf, J. C. (1988). Litigating personnel measurement policy. *Journal of Vocational Behavior, 33,* 235–271.

Society for Industrial and Organizational Psychology. (1987). *Principles for the validation and use of personnel selection procedures* (3rd ed.). College Park, MD: Author.

Sparrow, M., Moore, M. H., & Kennedy, D. M. (1990). *Beyond 911: A new era for policing.* New York: Basic Books.

Topping, R. (1995, November 17). Will "the test" pass the test? *Newsday,* pp. A5, A28.

United States v. Nassau County, CV 77 1881 Consent order (E.D.N.Y.) [Docket No. 354] (1995a, September 22).

United States v. Nassau County, CV 77 1881 (E.D.N.Y.) [Docket No. 365] (1995b, September 22).

Widgor, A. K., & Hartigan, J. A. (Eds.). (1988). *Interim report: Within-group scoring of the General Aptitude Test Battery.* Washington, DC: National Academy Press.

Zedeck, S., Cascio, W. F., Goldstein, I. L., & Outtz, J. (1996). Sliding bands: An alternative to top-down selection. In R. S. Barrett (Ed.), *Fair employment strategies in human resource management* (pp. 222–234). Westport, CT: Quorum Books.

Zelnick, R. (1996). *Back fire: A reporter's look at affirmative action.* Washington, DC: Regnery Publishing.