9-17-2010

Junk science, test validity, and the Uniform Guidelines for Personnel Selection Procedures: The case of Melendez v. Illinois Bell

Fred B. Bryant
Loyola University Chicago, fbryant@luc.edu

Elaine K.B Siegel
Hager & Siegel, P.C.

Optimal Data Analysis
2010, Vol. 1, Release 1 (September 17, 2010)

2155-0182/10/$3.00

Junk Science, Test Validity, and the Uniform Guidelines for Personnel Selection Procedures: The Case of *Melendez v. Illinois Bell*

Fred B. Bryant, Ph.D. and Elaine K.B. Siegel
Loyola University Chicago Hager & Siegel, P.C.

This paper stems from a recent federal court case in which a standardized test of cognitive ability developed by AT&T, the Basic Scholastic Aptitude Test (BSAT), was ruled invalid and discriminatory for use in hiring Latinos. Within the context of the BSAT, we discuss spurious statistical arguments advanced by the defense, exploiting certain language in the current Uniform Guidelines for evaluating the fairness and validity of personnel selection tests. These issues include: (a) how to avoid capitalizing on chance; (b) what constitutes "a measure" of job performance; (c) how to judge the meaningfulness of group differences in performance measures; and (d) how to combine data from different sex, race, or ethnic subgroups when computing validity coefficients for the pooled, total sample. Pursuant to the Uniform Guidelines' standard for unfairness, when one ethnic group scores higher on an employment test, the test is deemed "unfair" if this difference is not reflected in a measure of job performance. Although studies validating selection instruments often survive the unfairness test, such data are vulnerable to bias and manipulation, if appropriate statistical procedures are not used. We consider both the benefits (greater clarity and precision) and the potential costs (loss of legal precedent) of revising the Uniform Guidelines to address these issues. We further discuss legal procedures to limit "junk science" in the courtroom, and the need to reevaluate validity generalization in light of Simpson's "false correlation" paradox.

The purpose of this paper is to share our insights from a recent federal court case, which we refer to as *Melendez*, involving a claim of employment discrimination in personnel selection, *Melendez v. Illinois Bell Telephone Company*, No. 90 C 5020 (N.D. Ill. Sept. 16, 1994),

2155-0182/10/$3.00

aff'd, 79 F.3d 661 (7th Cir. 1996).[1] These insights arise from certain defenses advanced by the employer, in which dubious statistical procedures were justified by language from current federal guidelines for validating personnel selection tests, the Uniform Guidelines for Employee Selection Procedures, promulgated jointly by the United States Equal Employment Opportunity Commission and the United States Departments of Labor, Justice, and the Treasury [43 Fed. Reg. 38,290 (August 25, 1978); EEOC, 29 CFR Part 1607]. We refer to these as the Uniform Guidelines.

After providing some background to the particular legal case involved, we describe the original validation studies that formed the heart of the litigation, and present research evidence which was the main point of contention at trial. After summarizing the evidence against the validity of the personnel selection test in question—the Basic Scholastic Aptitude Test (BSAT)—we highlight some apparent ambiguities in the Uniform Guidelines. Comparable ambiguities exist in both the Standards for Educational and Psychological Testing[2] and in the Society for Industrial and Organizational Psychology's Principles for the Validation and Use of Personnel Selection Procedures.[3] Ironically, although the Uniform Guidelines are intended to promote equality of employment opportunity regardless of race, religion, and gender, they do not expressly prohibit the use of certain research practices that produce spurious artifacts, and which actually perpetuate discrimination in the workplace.

In this paper we share our observations with professionals within the psychological testing, statistical analysis, human resources and legal communities; discuss the application of Uniform Guidelines in maintaining consistency vis-à-vis professional standards; and conclude by recommending a reevaluation of the procedure of validity generalization in light of Simpson's "false correlation" paradox (i.e., paradoxical confounding).

Historical Context

What was this trial all about? Plaintiff Carmelo Melendez claimed he was denied equal employment opportunity in applying for a job with defendant Illinois Bell Telephone Company. Mr. Melendez was born and raised in Puerto Rico, and moved to East Chicago in the middle of his grade school years. Though he spoke no English, Mr. Melendez was placed in a monolingual English classroom. A straight-A student in Puerto Rico, in the United States he got F's. By struggling hard, he learned English, taught himself the skills he needed to advance, and raised his grades until, by the time he graduated from high school, he was earning B's.

It was then, however, that Mr. Melendez first encountered an obstacle that he could not overcome, and that he would confront throughout his adult life: standardized ability tests. He performed miserably on the SAT, and could not attend college. He decided to apply for an entry-level position in metallurgy at the local steel mill. He failed the standardized entry examination, however. Yet another standardized test kept him out of the military.

Mr. Melendez persevered, and eventually got his college degree. He also became a certified x-ray technician, and he eventually worked for the federal Civil Rights Commission. He went on to become the host of a Chicago-area television talkshow. Then, in 1988, he applied for a job as Assistant Manager of Urban Affairs for Illinois Bell.

The job description called for a person who could interface with the local Latino community, to assess emerging urban trends for use in marketing telecommunications services. The successful applicant should be able to interact with community leaders and residents, and to communicate effectively in a bilingual setting, orally and in writing.

Illinois Bell required all external applicants for its first-level management jobs to surmount three separate pass-fail hurdles. Applicants had to have a college diploma, graduating

2155-0182/10/$3.00

in the top half of the class. Applicants had to pass a structured, standardized interview, demonstrating a sufficient level of leadership. Finally, applicants had to take the standardized Basic Scholastic Aptitude Test (BSAT), scoring at or above a raw pass-fail cutoff score of 196. This cognitive ability test was the central focus of the court case.

The BSAT is a standardized paper-and-pencil test, purporting to assess verbal and quantitative ability, much like the SAT. It also includes questions designed to tap the ability to follow directions, in which one must indicate answers while listening to a tape-recording which contains complex, conflicting instructions. Each subsection of the test is timed, or "speeded," and the entire test takes about one hour.

Despite his college degree and his success on the leadership interview, Mr. Melendez failed the BSAT. He grew depressed and despondent, and became estranged from his family for more than a year. Not long after his rejection by Illinois Bell, however, Melendez won a position with the federal government. He has performed successfully there ever since, and has risen to a position of authority.

Based on his experience, Mr. Melendez believed that the BSAT was unfair because it was not job-related. He saw no connection between the skills required to do well on the job of Assistant Urban Affairs Manager, and the skills required to pass the BSAT. To right the wrong, he filed suit against Illinois Bell for employment discrimination.

Adverse Impact of the BSAT

Before turning to the evidence concerning test validity, we first consider the BSAT's impact on applicants of different ethnicity (i.e., the BSAT pass-fail rates for different racial or ethnic groups). Table 1 presents pass-fail rates for whites, African-Americans and Latinos on the BSAT separately for two time periods: 1979 and 1987-88. The 1979 statistics are for 591 managerial applicants, and are taken directly from the original AT&T validation report: in 1979, about 3 in 4 whites passed the test, versus 1 in 5 African-Americans, and 1 in 2 Latinos.[4]

Table 1: Rates of Success and Failure on the BSAT for Different Racial Groups

		Racial Group					
		White		Black		Latino	
Time Period		P	F	P	F	P	F
1979	*n*	265	79	42	151	25	29
	%	77	23	22	78	47	53
1987-88	*n*	344	51	83	62	50	44
	%	87	13	57	43	53	47

Between-Group Pairwise Comparisons via Fisher's Exact Test

	W79	W87	B79	B87	L79
W87	.000459				
B79	.000001	.000001			
B87	.000018	.000001	.000001		
L79	.000010	.000001	.000827	.21	
L87	.000014	.000001	.000001	.60	.50

Note: Pairwise comparisons were performed using two-tailed Fisher's exact test computed using ODA software.[5] Row and column headings indicate both ethnic class (W= white, B=Black, L=Latino) and time period (79=1979, 87=1987-88). Tabled for each unique combination of row and column is the *p*-value (six significant digits) for the exact test comparing pass/fail rates of the corresponding samples. *P*-values indicated in red are statistically significant at experimentwise $p<0.05$ based on an appropriate Bonferroni criterion (see discussion in paper: $p<.05/1115$, or $p<0.000046$); *p*-values indicated in blue are statistically significant at the generalized criterion (per-comparison $p<0.05$); *p*-values indicated in black are not significant.[5]

The 1987-88 pass-fail statistics are from Illinois Bell's records, from a sample of 634

2155-0182/10/$3.00

applicants for first-level management positions. During the 1987-88 period, most whites—nearly 9 in 10—passed the test, versus 6 in 10 African-Americans and 5 in 10 Latinos.

To evaluate these pass-fail rates, there is a guideline for judging the impact of an employment test on different ethnic groups. This rule-of-thumb is known as the "four-fifths rule." According to this guideline, a test has an *adverse impact* on an ethnic group whose pass rate is less than four-fifths the rate of the group with the highest test pass-rate: "A selection rate for any race, sex, or ethnic group which is less than four-fifths (4/5) (or eighty percent) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact, while a greater than four-fifths rate will generally not be regarded by Federal enforcement agencies as evidence of adverse impact" (Uniform Guidelines, §1607.4.D). The Uniform Guidelines define "adverse impact" as: "A substantially different rate of selection in hiring, promotion, or other employment decision which works to the disadvantage of members of a race, sex, or ethnic group" (Uniform Guidelines, §1607.16.B).

In 1979, for example, whites had the highest pass-rate on the BSAT, at 77% (see Table 1). The BSAT, then, had an adverse impact on any group in 1979 whose BSAT pass-rate falls below four-fifths of 77% (or below 61.6%). The 1979 pass-rates for African-Americans (22%) and Latinos (47%) are clearly lower than the four-fifths mark of 61.6%.

For the 1987-88 period, under the Uniform Guidelines' four-fifths rule, the BSAT had an adverse impact on any group whose pass-rate falls below four-fifths of the white pass rate of 87% (or below 69.6%). Because pass rates for African-Americans (57%) and Latinos (53%) are below this four-fifths mark of 69.6%, the BSAT had an adverse impact on both of these groups during 1987-88, according to the Uniform Guidelines' standard.

This evidence of strong and consistent adverse impact makes test validity even more vital. Rejecting such a large number of minority applicants might be defensible, if the test accurately predicted important on-the-job performance. For example, imagine using a valid test of visual acuity to select fighter-pilots; if minority applicants have worse eyesight than majority applicants, then so be it. It is an entirely different matter, however, if the test has nothing to do with on-the-job performance. If minorities do not actually have worse eyesight, then the test unfairly denies them equal employment opportunity. In the case of the BSAT, the evidence for test validity is particularly critical, given the unequivocal adverse impact on minorities. In the words of the Uniform Guidelines: "Reliance upon a selection procedure which is significantly related to a criterion measure, but which is based upon a study involving a large number of subjects and has a low correlation coefficient will be subject to close review if it has a large adverse impact..." (Uniform Guidelines, §1607.14(B)(6).

BSAT Validation Studies

Two validation studies of the BSAT formed the heart of the litigation, and the trial gravitated around certain research evidence from these studies. In the late 1970s, AT&T industrial/organizational psychologists developed the BSAT, using test components originally written by the Educational Testing Service (ETS), which also developed the SAT, LSAT, GRE, and other cognitive ability tests. One of the AT&T psychologists drafted the final research report containing two validation studies, which assessed the relationship between BSAT scores and job performance. These studies purported to evaluate the BSAT's predictive validity, i.e., its ability to predict subsequent on-the-job performance. Illinois Bell relied on these validation studies in using the BSAT to screen its job applicants.

2155-0182/10/$3.00

The first of the two validation studies, referred to as the *Preliminary Study*, focused on entry-level managers already hired at 8 different company locations throughout the country. This Preliminary Study included 229 managers who had earlier taken a large battery of standardized tests, including the School and College Ability Test (SCAT) and the predecessor of the BSAT, the Bell System Qualification Test (BSQT). One year after these applicants were hired their job performance was evaluated by their supervisors, who rated each applicant's job performance using a set of 13 criterion measures, developed through a job analysis of management positions, including ratings of skills in planning, decision making, oral and written communications, leadership, resistance to stress, interpersonal awareness, and a global rating of overall job performance. The test developers then selected a subset of verbal and math items based on correlations with supervisor ratings, and these items became the BSAT. Researchers then examined the relationship between test score and rating of overall job performance to establish a pass-fail cut-score for the test, which was implemented throughout AT&T companies.

The second validation study, referred to as the *Followup Study*, focused on 286 job applicants who were applying for entry-level management positions in 11 different AT&T company locations. Applicants selected for participation were given the BSAT (using the pass-fail cut-score determined in the Preliminary Study), and then one year later, their supervisors were asked to rate each employee on a set of 15 performance criteria. As in the Preliminary Study, researchers examined the correlation between test scores and performance ratings, trying to cross-validate the findings from the Preliminary Study. Thus, both validation studies concern the predictive validity of the test, that is, whether the test accurately predicts job performance and is therefore job-related.

Validity Evidence for the BSAT

What evidence is there concerning the predictive validity of the BSAT? The primary validity evidence in the validation studies consists of Pearson product-moment correlation coefficients relating applicants' test scores to supervisors' performance ratings.

Preliminary Study. Turning first to Table 2, note that the Preliminary Study reports no figures for Latinos. Instead, for African-Americans and whites separately and for the pooled data set, it reports correlations between BSAT scores and each of the 13 performance ratings. Note that the BSAT shows a statistically significant correlation with ratings of overall job performance for the total sample, $r(151)=0.38$, $p<0.00001$. For whites, however, only 4 of the 13 criterion measures show a statistically significant ($p<0.05$) relationship with BSAT score. Indeed, BSAT scores had no significant relationship with ratings of overall job performance for whites. Averaging across all correlations for whites (mean $r=0.128$, $p<0.08$), the BSAT predicts about 2% of the variance in whites' performance ratings. This represents a Hedges corrected effect-size of 0.26, equivalent to an experimental effect in which the treatment group scores about one-quarter of a standard deviation above the control group.

Also note that, for African-Americans, 7 of the 13 performance ratings (including overall job performance) show a statistically significant relationship with BSAT score. Averaging across all correlations (mean $r=0.314$, $p<0.006$), the BSAT explains about 10% of the variance in African-Americans' performance ratings (Hedges corrected g=0.65). Considered together, this evidence from the Preliminary Study suggests that the BSAT is largely invalid for use with whites, but has marginal validity for use with African-Americans. We return later to the first column of Table 2, giving validity coefficients for the total group.

2155-0182/10/$3.00

Table 2: Preliminary Study Correlations Between BSAT Score and Job Performance Ratings for Different Groups

	Groups		
Job Skills	Total *n*=153	White *n*=94	Black *n*=39
Organizing and Planning	.28*	.09	.34*
Decision Making	.30*	.20*	.27
Decisiveness	.39*	.25*	.36*
Oral Communications	.23*	.08	.43*
Written Communications	.28*	.21*	.26
Leadership	.36*	.02	.54*
Interpersonal Awareness	.25*	.09	.30*
Behavior Flexibility	.20*	.04	.20
Fact Finding	.38*	.29*	.24
Resistance to Stress	.21*	.11	.18
Energy Management	.15	.04	.08
Potential	.42*	.11	.42*
Overall Job Performance	.38*	.13	.46*

Note: Adapted from Tables 4 and 8 of the original validation report.[4] An asterisk (*) indicates $p<0.05$ at the generalized (per-comparison) criterion.[5] *N* for the total sample is greater than the sum of the *n*s for the white and black groups because the Preliminary Study included 16 Hispanics and 4 "other minorities" whose data were pooled in the analysis of the total sample. Discussed further ahead in the paper, the "false correlation paradox" (paradoxical confounding) is present when an index for pooled samples lies outside the range of index values for individual samples considered separately (indicated in red).

Followup Study. Table 3 gives validity coefficients for the Followup Study. Again the BSAT shows a significant correlation with ratings of overall job performance for total sample, $r(284)=0.21$, $p<0.001$. For whites, 4 of 15 performance ratings show a significant relationship with BSAT score: averaging coefficients (mean $r=0.077$, $p>0.19$), the BSAT predicts about 2% of the variance in whites' performance ratings (corrected $g=0.19$). For African-Americans, 8 of 15 validity coefficients are significant: averaging coefficients (mean $r=0.215$, $p<0.01$), the BSAT predicts about 6% of the variance in African-Americans' performance ratings (corrected $g=0.44$). BSAT score was significantly related to ratings of overall job performance for both whites and African-Americans, though these effect sizes again were relatively small.

The fourth column in Table 3 reports the only direct empirical evidence available concerning the validity of the BSAT for use in hiring Latinos. Only one of the 15 validity coefficients was significantly different from zero for Latinos ($r=0.24$, $p<0.05$, one-tailed) for Latinos. The sole significant coefficient (for coordination) was reported as nonsignificant in the original validation study. Essentially, this means that the BSAT does no better than chance in predicting how Latinos will perform on the job (mean $r=0.093$, $p>0.32$, corrected $g=0.21$).

In relation to the present case, this is the single most relevant piece of validity evidence in the entire report. *Plainly, these data do not support the validity of using the BSAT to hire Latinos.*

Inflation of Apparent Validity Vis-à-Vis Extensive Analysis: The "Trolling" Problem

It would be one matter if the coefficients were the only analyses in the validation studies. If this were the case, then there would be 49 tests of statistical hypotheses in the Preliminary Study (Table 3) and 60 tests in the Followup Study (Table 4), for a total of 109 tests.

2155-0182/10/$3.00

Table 3: Followup Study Correlations Between BSAT Score and Job Performance Ratings for Different Groups

	Groups			
Job Skills	Total (*n*=286)	White (*n*=147)	Black (*n*=76)	Latino (*n*=57)
Organizing and Planning	0.17*	0.08	0.19	0.15
Decision Making	0.18*	-0.12	0.21*	-0.08
Oral Communications	0.17*	0.10	0.26*	0.01
Written Communications	0.28*	0.18*	0.44*	0.10
General Administration	0.11*	0.09	0.22*	0.07
Supervision	0.01	0.02	0.10	0.09
Coordination	0.19*	0.01	0.30*	0.24*
Behavior Flexibility	0.10*	0.03	0.20	0.08
Fact Finding	0.25*	0.10	0.33*	0.18
Problem Solving	0.22*	0.17*	0.25*	0.08
Resistance to Stress	0.05	0.06	0.05	0.05
Ability to Learn and Develop	0.16*	0.05	0.17	0.10
Tolerance of Ambiguity	0.12*	0.08	0.17	0.07
Management Potential	0.16*	0.16*	0.08	0.12
Overall Job Performance	0.21*	0.14*	0.26*	0.14

Note: Adapted from Table 18 of the original validation report .[4] *N* for the total sample is greater than the sum of the *n*s for the three subgroups because the Followup Study included six Asians whose data were pooled for total sample analysis. An asterisk (*) indicates $p<0.05$ at the generalized (per-comparison) criterion. The coefficient indicated in red was reported as being nonsignificant in the original validation report, but is actually statistically significant at the generalized criterion ($p<0.05$, one-tailed).

Tallying across the entire validation report, however, reveals that more than a thousand statistical tests were performed—all using the $p<0.05$ level of statistical significance. Of those 1000 tests, 50 would be expected simply by chance alone to be statistically significant at per-comparison $p< 0.05$, although exactly which effects are attributable to chance cannot be known. The validity evidence is thus inflated, as the excessive statistical testing adds a substantial number of chance correlations to the true correlations. Accordingly, well-known procedures for controlling the experimentwise Type I error-rate should be used.[5] For example, among the most commonly employed methods for reducing the number of "false-positive" results when conducting numerous statistical tests is the so-called "Bonferroni adjustment, in which an adjusted *p*-value is obtained by dividing the desired alpha-level by the number of *p*-values examined. For the BSAT validation report, a Bonferroni-adjusted *p*-value would be roughly .05/1100, or $p<0.00005$. This is the cost for undertaking vast numbers of analyses indiscriminately, when analyses can and should be more clearly focused.[5,6]

2155-0182/10/$3.00

Table 4: Followup Study Means and Standard Deviations for the 15 Job Performance Ratings, and for BSAT Score for Whites (*n*=147) and Latinos (*n*=57)

	Whites		Latinos	
Job Skills	Mean	sd	Mean	sd
Organizing and Planning	5.22	1.13	4.99	1.04
Decision Making	5.15	0.93	4.93	0.84
Oral Communications	5.31	1.15	4.88	1.18
Written Communications	5.24	1.16	4.82	1.23
General Administration	5.12	1.06	4.68	0.87
Supervision	4.98	1.23	4.92	1.32
Coordination	5.39	1.00	4.85	0.90
Behavior Flexibility	5.25	1.17	4.83	1.08
Fact Finding	5.38	1.11	4.88	1.06
Problem Solving	5.18	1.03	4.86	1.10
Resistance to Stress	5.22	1.11	5.25	0.97
Ability to Learn and Develop	5.71	1.01	5.41	1.20
Tolerance of Ambiguity	5.08	1.13	4.81	0.86
Management Potential	6.02	1.93	6.65	2.08
Overall Job Performance	5.35	1.06	5.11	1.08
BSAT score	218.62	13.89	209.78	15.49

Note: Adapted from Tables 14 and 17 of the original validation report.[4] Scores on the 7-point rating scales have been reversed so that high scores reflect better ratings. Means indicated in red differ from the mean for whites with $p<0.05$ by Tukey's Honest Significant Difference multiple range test. These statistically significant group differences were found when following up significant *F*-values from initial one-way analyses of variance with white, Latino, and African-American groups.

In the *Melendez* case, we took the "middle-ground" approach of adjusting the criterion to $p<0.05$ in the validation studies. This reduces spurious effects (Type I errors), without unduly increasing false no-difference conclusions (Type II errors) due to low statistical power. Evaluated at this criterion, *there are no significant validity coefficients in the Followup Study*.

Illinois Bell defended its inflationary statistical procedures with a statement in the Uniform Guidelines that one should usually use the $p<0.05$ level in establishing statistical significance: "...Generally, a selection procedure is considered related to the criterion, for the purposes of these guidelines, when the relationship between performance on the procedure and per-

2155-0182/10/$3.00

formance on the criterion measure is statistically significant at the $p<0.05$ level of significance" (Uniform Guidelines, §1607.14.B(5)). The Uniform Guidelines nonetheless require the use of "professionally acceptable statistical procedures" in computing validity coefficients (Uniform Guidelines, §1607.14.B(5)), and also caution users to avoid using procedures that capitalize on chance: "*Overstatement of validity findings*. Users should avoid reliance upon techniques which tend to overestimate validity findings as a result of capitalization on chance unless an appropriate safeguard is taken. Reliance upon a few selection procedures or criteria of successful job performance when many selection procedures or criteria of performance have been studied, or the use of optimal statistical weights for selection procedures computed in one sample, are techniques which tend to inflate validity estimates as a result of chance. Use of a large sample is one safeguard; cross-validation another." (Uniform Guidelines, §1607.14.B.(7).

Clearly, performing 1100 statistical tests at the $p<0.05$ level is a procedure that capitalizes on chance. Under the Guidelines, an adjustment to the alpha-level is in order, minimally one such as using $p<0.01$. To reduce jury confusion over these technical issues, the Uniform Guidelines should include specific recommenddations (e.g., Bonferroni adjustments) for reducing Type I error when a large number of statistical tests have been conducted.

Filling the Validity Gap with Junk Science: Reinventing Statistics

Through the above evidence, plaintiff demonstrated that the BSAT had, at most, negligible validity for white applicants, and no validity for Latino applicants. And how did Illinois Bell respond to plaintiff's showing? Illinois Bell's expert witness, an organizational psychologist, asserted that if the BSAT truly had a nonsignificant (i.e., zero) statistical relationship with job performance for Latinos, then half of the validity coefficients for Latinos should have been positive, and half negative. In other words, if the true value of the correlation in the population is zero, then there should be just as many positive validity coefficients as negative. He noted, however, that 14 of the 15 coefficients for Latinos in the Followup Study were positive (if not statistically significant). He then calculated the binomial probability of obtaining 14 positive coefficients and 1 negative, given a 0.50 probability for obtaining either sign (i.e., $z=3.30$, $p<0.0005$). From this scenario, he deduced that, despite the complete lack of any correlation in the AT&T validation study, the BSAT was nonetheless valid for Latinos—and at a highly significant *p*-value!

By pitting one expert's statistical analysis against the other's, this form of "junk science" has great potential to confuse the jury. To clarify the issue for the layperson, what is needed is a logical, easy-to-follow explanation of the difference between the two opposing views of the same data. However, this is not always easily developed.

In the *Melendez* case, we explained the statistical issue in commonsense terms by using an archery analogy. Testing the validity of the BSAT is like an archery contest. An archer fires 15 arrows at a target; to determine his proficiency, we count how many arrows hit the target. Using the BSAT to predict the 15 performance criteria for Latinos, we count how many times it shows a statistically significant relationship between test score and job performance. Table 3 shows that for Latinos, all 15 arrows missed the mark. By the rules of the game, the archer does not score, and the BSAT is off target (and invalid).

By Illinois Bell's logic, however, 14 of the 15 arrows flew in the target's general direction (i.e., 14 of the 15 validity coefficients were positive) and only 1 arrow flew in the opposite direction (i.e., there was only one negative validity coefficient), and so therefore the archer was a success (and the BSAT is valid for Latinos because only one of its validity coefficients was

2155-0182/10/$3.00

negative). This is fallacious. At issue is the *magnitude* of the validity coefficients in the positive direction, not just whether the signs of these coefficients are positive or negative. For the BSAT, the magnitudes were insufficient to establish a statistically significant relationship. As the Seventh Circuit ruled on appeal, there was "strong evidence of the BSAT's inability to predict job performance," which supported the trial court's finding that "the BSAT's discriminatory impact was unjustified by Illinois Bell's legitimate business needs" (79 F.3d at 669). That is, the BSAT explains too little variance in performance ratings to be considered valid for use in hiring Latinos. If the BSAT does not provide useful, job-related information, then its use cannot be justified, given the strong evidence of its adverse impact.

The Admissibility of "Junk Science" in the Courtroom

Illinois Bell's spurious defense, that its test is "valid" because of its positive (though not statistically significant) correlations with performance ratings, exemplifies the dangers of "junk science" in the courtroom. As the U.S. Supreme Court has cautioned: "Expert evidence can be both powerful and quite misleading because of the difficulty in evaluating it." (*Daubert*, 509 U.S. at 595 (quoting Weinstein, 1992)). Due to defendant's discovery abuse, Melendez was able to bar, altogether, the testimony of the company's expert witness. More typically, dubious science is precluded through a ruling by the trial court that the information is inadmissible under the Federal Rules of Evidence.

Expert testimony is specifically governed by Federal Rule of Evidence 702, which establishes ground rules for admitting expert testimony: "If scientific, technical, or other specialized knowledge will assist the trier of fact to understand the evidence or to determine a fact in issue, a witness qualified as an expert by knowledge, skill, experience, training, or education, may testify thereto in the form of an opinion or otherwise" (Fed. R. Evid. 702). As interpreted in the landmark *Daubert* decision, Rule 702 allows expert testimony when it is both relevant and scientifically reliable. In *Daubert* the Court appointed the trial judge as the "gatekeeper" of expert testimony, asserting: "[t]his entails a preliminary assessment of whether the reasoning or methodology underlying the testimony is scientifically valid and of whether that reasoning or methodology properly can be applied to the facts in issue." (*Daubert*, 509 U.S. at 592-593). The Court went on to explain: "The inquiry envisioned by Rule 702 is, we emphasize, a flexible one. Its overarching subject is the scientific validity—and thus the evidentiary relevance and reliability—of the principles that underlie a proposed submission. The focus, of course, must be solely on principles and methodology, not on the conclusions that they generate" (*Daubert*, 509 U.S. at 594-595).

More recently, the U.S. Supreme Court held unanimously that a trial court's decision to admit or exclude expert evidence should be accorded great deference (*Joiner*, 118 S.Ct. 512). Noting that trial judges typically are not scientists, Supreme Court Justice Stephen Breyer encouraged judges to take the initiative to clarify scientific issues (*Joiner*, 118 S.Ct. 512, 520-521 (Breyer, J., concurring)). They may, for example, utilize their authority to appoint their own experts, or use pretrial hearings to explore the issues. The *Daubert* Court explains that the goal is a middle ground, between "a 'free-for-all' in which befuddled juries are confounded by absurd and irrational pseudoscientific assertions", and "a stifling and repressive scientific orthodoxy" (*Daubert*, 509 U.S. at 595-596). The Court recalled the differences between scientific inquiry and the law, emphasizing that Federal Rules of Evidence are "designed not for the exhaustive search for cosmic understanding but for the particularized resolution of legal disputes" (*Daubert*, 509 U.S. at 597).

2155-0182/10/$3.00

The Concept of Test "Fairness"

Besides adverse impact and validity, another critical concept in judging whether or not a test in discriminatory is test "fairness." Although researchers have suggested numerous definitional frameworks and statistical models of test fairness[7-12], two approaches are often used in litigation to define "unfairness," and to determine whether a test is "unfair."

Anne Cleary[13] pioneered one of these definitions at the Educational Testing Service. According to Cleary's model, a test is considered "unfair" when it predicts performance differently for different ethnic groups. This differential prediction is detected in the form of statistically significant differences between groups in the slopes and in the intercepts of the regression lines relating test scores to performance. Thus, a test is considered "fair" when there are no significant differences in errors of prediction between groups, using a common regression line. Ironically, by a strict application of Cleary's definition, an invalid test could be deemed "fair." It would not be unfair, for example, to use a coin-flip to hire job applicants, because this selection procedure does not predict performance better for one ethnic group than for another. It is equally invalid for both groups.

Another definition of "unfairness" prominent in the courts is that used in the Uniform Guidelines, under which a test is "unfair" when: "...members of one race, sex, or ethnic group characteristically obtain lower scores on a selection procedure than members of another group, and the differences in scores are not reflected in differences in a measure of job performance..." (Uniform Guidelines, §1607.14.B(8)(a)).

In practice, one determines whether a test is "unfair" by comparing group means on the test, then looking for comparable mean-differences in group performance ratings. If one group scores higher on the test, it must also do better on the job. Stated differently, a test is "unfair" if it denies job opportunities to a group whose actual job performance is up to par.

Applying the Uniform Guidelines' definition of "unfairness" to the BSAT Followup Study, Latinos had significantly lower BSAT scores than whites, and passed the test at a significantly lower rate (77% *vs*. 47% in 1979; 87% *vs*. 53% in 1987-88; Table 1). In contrast, on 12 of the 15 performance criteria, Latino and white performance ratings did not differ significantly (Table 4). In other words, 80% of the performance measures (including overall job performance) failed to show lower scores for Latinos than whites. Considered together, this evidence shows that the BSAT is "unfair" to Latinos within the meaning of the Uniform Guidelines.

Twisting the Uniform Guidelines to Establish Test "Fairness"

In a spurious defense of the BSAT, Illinois Bell purported to rely on the Uniform Guidelines' definition of test unfairness. At trial the defense argued that the company adhered to the letter of the Uniform Guidelines, and advanced two lines of defense based on the Guidelines. Neither the law nor professional standards support these arguments.

What constitutes "a measure of job performance"? On cross-examination, the defense read to the jury the Uniform Guidelines' definition of test unfairness in Section 14.B(8)(a), and then asked:

Q: "Am I correct, Doctor, that this says that the differences in scores are not reflected in differences in a measure of job performance? Do you see that, Doctor?"

A: "Yes, I do."

Q: "And you have just testified that here there are three measures of job performance at which Whites score statistically higher than Hispanics, is that correct Doctor?"

A: "That's correct."

2155-0182/10/$3.00

Q: "So according to this definition which you have been relying on, there is not unfairness in this test, isn't that right, Doctor?"

The trial court struck this line of questioning. Illinois Bell's interpretation of the Uniform Guidelines' definition of "test unfairness" lacks any scientific or legal basis. While the term "measure" may signify either a single item, or a set of items measuring a single latent construct, this is no mere semantic quibble. What constitutes a "measure," in a given context, must be determined through appropriate legal and statistical analysis.

As a legal matter, Illinois Bell's interpretation of Section 14.B(8)(a) ignores its precise language. Through the use of the phrase "differences in *a* measure," the Uniform Guidelines plainly contemplate "a measure" as comprising more than one item. This conclusion is reinforced by the language of the definition of "unfairness" in the "Definitions" section of the Uniform Guidelines: "*Unfairness of selection procedure*. A condition in which members of one race, sex, or ethnic group characteristically obtain lower scores on a selection procedure than members of another group, and the differences are not reflected in differences in measures of job performance. See section 14.B.(7)" (Uniform Guidelines, §1607.16.V) [emphasis added]. The two definitions of "unfairness" must be read together, and thus do not support reliance on an isolated difference in measurement ("Definitions" section of the Uniform Guidelines mandates "[t]he following definitions *shall apply* throughout these guidelines" (Uniform Guidelines, §1607.16) [emphasis added]).

Illinois Bell's argument, moreover, would permit an employer to ignore the vast weight of unfavorable evidence, so long as any favorable evidence existed at all. Defendant's interpretation would render the unfairness standard meaningless. The term "measure" cannot be applied arbitrarily, but requires a fact-sensitive analysis.

In the *Melendez* case, we reanalyzed the correlations among the 15 performance ratings using both exploratory and confirmatory factor analysis.[14] We found that the 15 criteria are most accurately represented as a single, global measure of job performance. Statistically, the 15 ratings are sufficiently interrelated so that they comprise not 15 independent measures, but rather only one underlying measure. The separate performance ratings cannot properly be considered individually.

Factor analysis should be used routinely in deciding whether to employ single items or composite scales to measure job performance. This would preclude test developers from treating sets of unidimensional criterion measures as multiple single-item indicators, and then selecting and highlighting, as evidence of test "fairness," any criteria on which the majority group has a higher mean. Confirmatory factor analysis, not subjective preference, should answer the question: "what is a measure?"

Factor analytic methodology adheres to the Uniform Guidelines, which proscribe "...reliance upon techniques which tend to overestimate validity findings as a result of capitalization on chance.... Reliance upon a few... criteria of successful job performance when many... criteria of performance have been studied... tend[s] to inflate validity estimates as a result of chance." (Uniform Guidelines, §1607.14.B(7)).

By what criterion should one judge differences in group means? On cross-examination, the defense inquired where the unfairness standard in the Uniform Guidelines requires that group differences be *statistically significant*. The Uniform Guidelines do not authorize excursions into chance associations, but the unfairness standard does not explicitly require statistical significance as a decision criterion. It should be noted, however, that the "Documentation" requirements of the Uniform Guidelines mandate the *reporting* of methods of data analysis, as well as the magnitude, direction, and statistical significance of results. It expressly re-

2155-0182/10/$3.00

quires that "[s]tatements regarding the statistical significance of results should be made (essential)." (Uniform Guidelines, §1607.15.B(8)). This section of the Guidelines specifically refers to measures of central tendency (e.g., means) and studies of test fairness. Illinois Bell argued, in essence, that professional statistical standards may somehow be suspended in evaluating employment test data.

Abandoning professional standards is scientifically and legally untenable. The Uniform Guidelines are themselves founded on the standards of the psychological profession. The Uniform Guidelines, §1607.1.C, states: "These guidelines have been built upon court decisions, the previously issued guidelines of the agencies, and the practical experience of the agencies, as well as the standards of the psychological profession."

Test developers should always adhere to professional standards for drawing inferences from data. The Guidelines do not require researchers to clear the memory of their calculator between computations, but researchers typically do so as a matter of course. Nor can employers ignore the Guidelines' prohibition against reliance on chance (Uniform Guidelines, §1607.14. B(7)). And yet, that is precisely the result if one relies on apparent group differences that lack statistical significance.

Illusory "Fairness" and Artifactual "Validity"

Under the Uniform Guidelines' "unfairness" standard, if one ethnic group scores higher than another on an employment test, and this difference is not reflected in a measure of job performance, the test is deemed "unfair." The BSAT failed this standard. Despite great disparities in test scores, whites and Latinos performed on the job with substantially similar success.

Importantly, under the Uniform Guidelines, the mere fact that majorities outscore minorities on an examination, while securing more favorable performance evaluations, does not affirmatively establish that the test is "fair." It does not prove the positive, that the test is "fair" and "job related," but it does disprove one possible negative. The standard, that is, should not be understood as establishing an affirmative defense for employers. Evidence that a test is not "unfair" merely forestalls the inference of discrimination that arises in cases when the group that excels on the test, garnering the greater share of job opportunities, does not actually do the job appreciably better. To prove or disprove "fairness," the parties may introduce other evidence.

Ironically, the pattern of data contemplated by the Uniform Guidelines' unfairness standard may result in a serious distortion of the validity evidence. If the data from different ethnic groups are simply (and improperly) combined in a pooled analysis, the distribution of the data will typically create the illusion of a correlation between test scores and performance ratings. Scatterplotting the data, the group with higher test scores and performance ratings will tend to fall in the upper right quadrant of the scatterplot. The group with lower test scores and performance ratings will tend to fall in the lower left quadrant of the scatterplot. This pattern will create an apparent correlation between test scores and performance ratings, despite the lack of any true relationship, and it will inflate obtained validity coefficients for the total sample. This problem is a variation of a phenomenon known as *Simpson's paradox*.[15,16]

The following hypothetical example demonstrates how the "false correlation" paradox can occur. Imagine that you are in the middle of a job interview. The interview is going well, so you broach the topic of salary. "How much would I be paid?" "Well," replies the interviewer, "take off your shoes, and let's find out." Requesting an explanation, you are told that the company has found that shoe size is a valid predictor of a person's worth. The company routinely measures the size of job applicants' feet, and then uses the results of that

2155-0182/10/$3.00

measurement to determine salary. Still skeptical, you ask to see the validity evidence, and the interviewer hands you a copy of a table from a research document (see Table 5).

TABLE 5: Validating Shoe Size as a Predictor of Salary: Hypothetical Raw Data for Women and Men

Women	Occupation	Shoe Size	Annual Salary
Ann	secretary	3	$ 22,000
Beatrice	actress	4	$ 14,000
Carol	teacher	4	$ 30,000
Diane	librarian	5	$ 20,000
Edna	lab technician	5	$ 40,000
Florence	baby sitter	5	$ 10,000
Gwen	journalist	6	$ 28,000
Harriet	bank teller	6	$ 18,000
Iris	nurse	7	$ 32,000
Jacqueline	waitress	7	$ 16,000
	Mean :	5.2	$ 23,000
Men	**Occupation**	**Shoe Size**	**Annual Salary**
Al	salesman	8	$ 48,000
Bob	airline pilot	8	$ 62,000
Carl	chef	9	$ 50,000
Don	chemist	10	$ 55,000
Ed	executive	10	$ 70,000
Frank	mechanic	10	$ 40,000
Greg	plumber	11	$ 52,000
Harold	electrician	11	$ 59,000
Ian	detective	12	$ 45,000
John	architect	12	$ 65,000
	Mean	10.1	$ 54,600
Exact Test of Gender Difference:		$p<0.000001$	$p<0.000547$

This table presents raw (hypothetical) data for a sample of 10 men and 10 women, listing their first name, occupation, shoe size, and salary. Reported at the bottom of the data table are the results of exact nonparametric statistical analyses[5] comparing men's and women's mean shoe-size (predictor) and salary (criterion). Women have smaller feet than men, and have comparably smaller salaries. Therefore, by the Uniform Guidelines' unfairness standard, it is not "unfair" to men or to women to use shoe size to determine salary. Validity coefficients relating shoe size to salary, and scatterplots of shoe size and salary, are presented in Figure 1.

2155-0182/10/$3.00

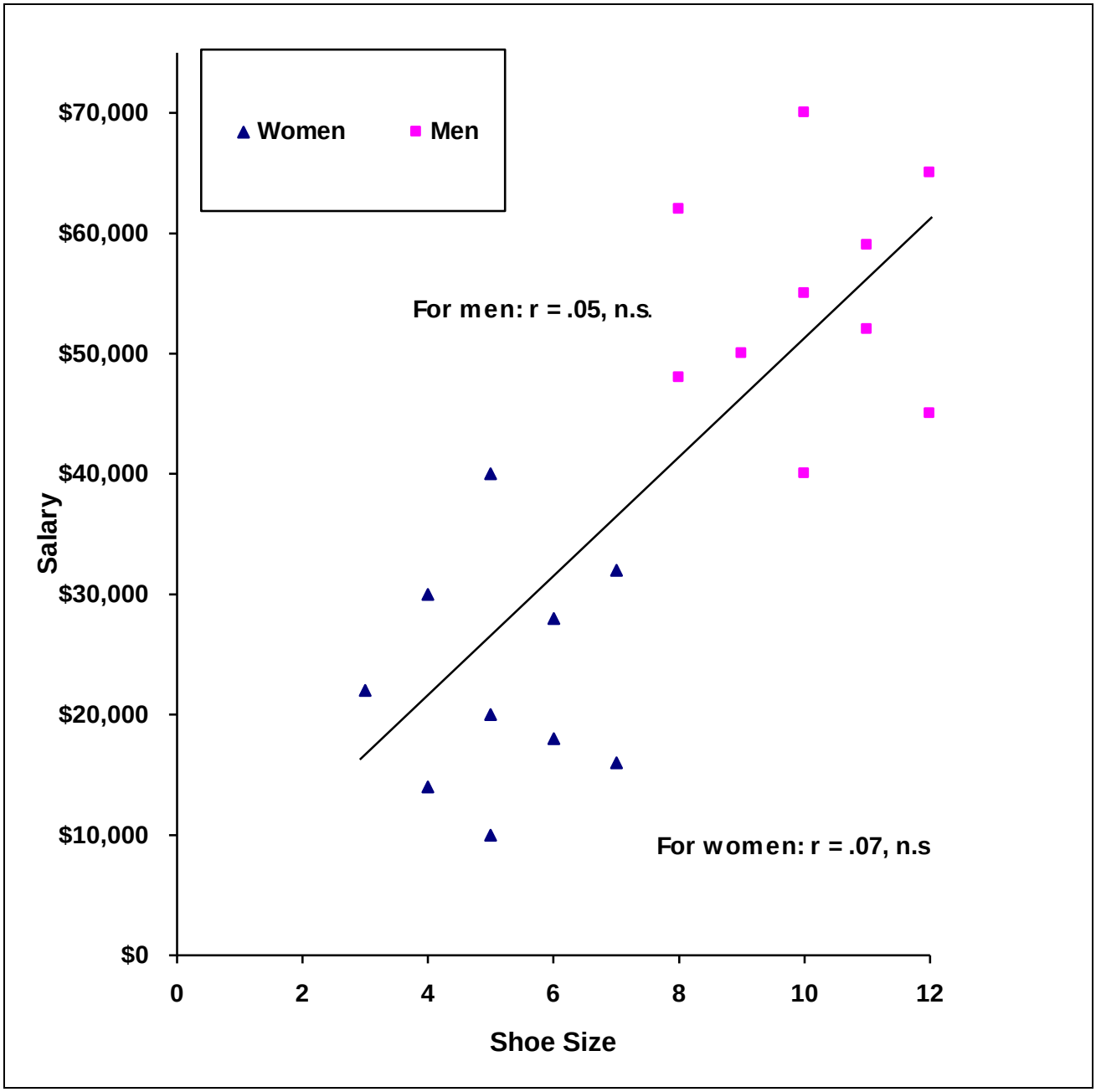


Figure 1: Correlating Shoe Size and Salary using Pooled Hypothetical Raw Data for Women and Men

Examination of validity coefficients for men and women reveals there is no linear relationship between shoe size and salary for either group: r=0.05 for men, r=0.07 for women, ps> 0.05. But, if men's and women's raw data are pooled, the men's data fall into the upper right-hand quadrant of the scatterplot, and the women's data fall into the lower left-hand quadrant (men score higher than women on predictor and criterion measures). When the correlation between shoe size and salary is computed for the total group of 20 subjects, r=0.78, p<0.001)! Based on this evidence and in accordance with the Uniform Guidelines, it is concluded that it is *both* fair and valid to use shoe size to determine salary.

2155-0182/10/$3.00

This hypothetical scenario is no more absurd than the BSAT validation work. In the Preliminary Study, for example, African-Americans had lower BSAT scores than whites, and they also had comparably lower performance ratings (thus the test does not meet the definition of unfairness, under the Uniform Guidelines' definition).

Figure 2 displays scatterplots of the group means on the BSAT and on overall job performance from the two validity studies. Clearly, these mean differences will inflate the apparent linearity of the relationship between BSAT and performance.

This inflation of correlations strikingly appears in the table of validity coefficients from the Preliminary Study (Table 2). Comparing the correlations of white, African-American, and total groups on the various performance measures, we find an anomalous pattern.

Consider the performance criterion of Decision Making. Its validity coefficient is r=0.20 for the group of 94 whites, and r=0.27 for the group of 39 African-Americans. For the Total Group, however, the r=0.30 correlation is higher than that for either subgroup. Similarly, the validity coefficients for Written Communications are r=0.21 for whites, r=0.26 for African-Americans, and r=0.28 for the Total Group; for Resistance to Stress, r=0.11 for whites, r= 0.18 for African-Americans, and r=0.21 for the Total Group; and for Energy, r=0.04 for whites, r=0.08 for African-Americans, and r=0.15 for the total group. Cases such as these, in which the correlations for the pooled group actually exceed the correlations found in each constituent subgroup, are a tell-tale sign of the "false correlation paradox," where in fact the "whole" is deceptively greater than the sum (or weighted average) of its parts.[16]

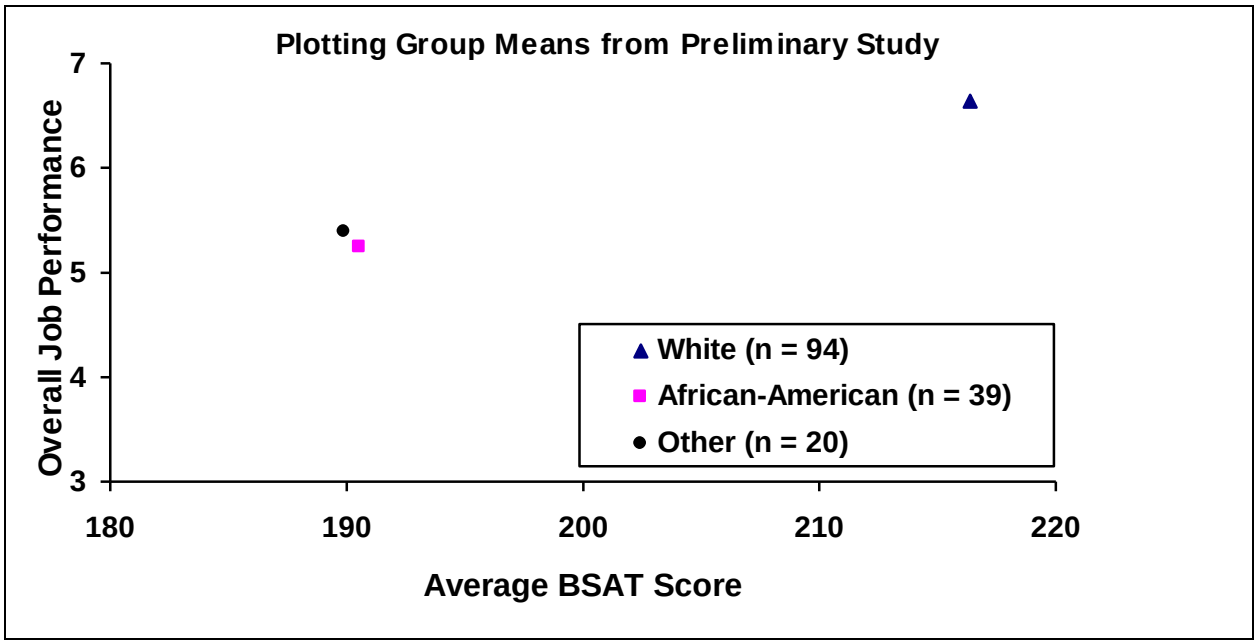
This technical problem is particularly critical because Illinois Bell rested its claim that the test was valid largely based on one number—one validity coefficient: the correlation between BSAT score and the rating of overall job performance, for the *Total Group* in the Preliminary Study. That coefficient is r=0.38, significant for the total sample of 153 subjects at $p<$ 0.00001 (see Table 2).

A *possible* methodology for circumventing such *paradoxical confounding* (the technical terminology for the "false-correlation problem") is to remove mean differences on the x- and y-variables before combining the data: for example, standardizing the x- and y-scores separately for each group using a *z*-score transformation maps the data into the same metric.[16] How does this work in the shoe size example? After transforming subjects' raw data to *z*-scores separately within the male and female samples, and subjecting these standardized data to correlation analysis, yields results given in Figure 3. When properly analyzed, the correlation between shoesize and salary is r=0.05 for men, r=0.07 for women, and r=0.06 for the total group.

This cure for Simpson's paradox (normatively standardizing separately by sample) only works if the true relationship between x and y is consistent across the multiple samples.[16] For example, if x and y are perfectly *positively* correlated in sample A and perfectly *negatively* correlated in sample B, normatively standardizing the data separately by sample and then combining them will yield a correlation coefficient of zero. Thus, it is necessary to verify homogeneity of covariance between x and y across samples before standardizing and pooling the data.[16-18]

Fortunately, instances of reverse validity rarely appear in the personnel selection literature.[19] Indeed, some proponents of validity generalization have even argued against the notion of differential validity altogether, though the BSAT data clearly show stronger evidence of validity for African-Americans than for Latinos or whites.[10] Thus, when analyzing the total sample, it should be routine practice before pooling data to normatively standardize separately within groups (after first verifying between-group equivalence of covariance matrices).

2155-0182/10/$3.00



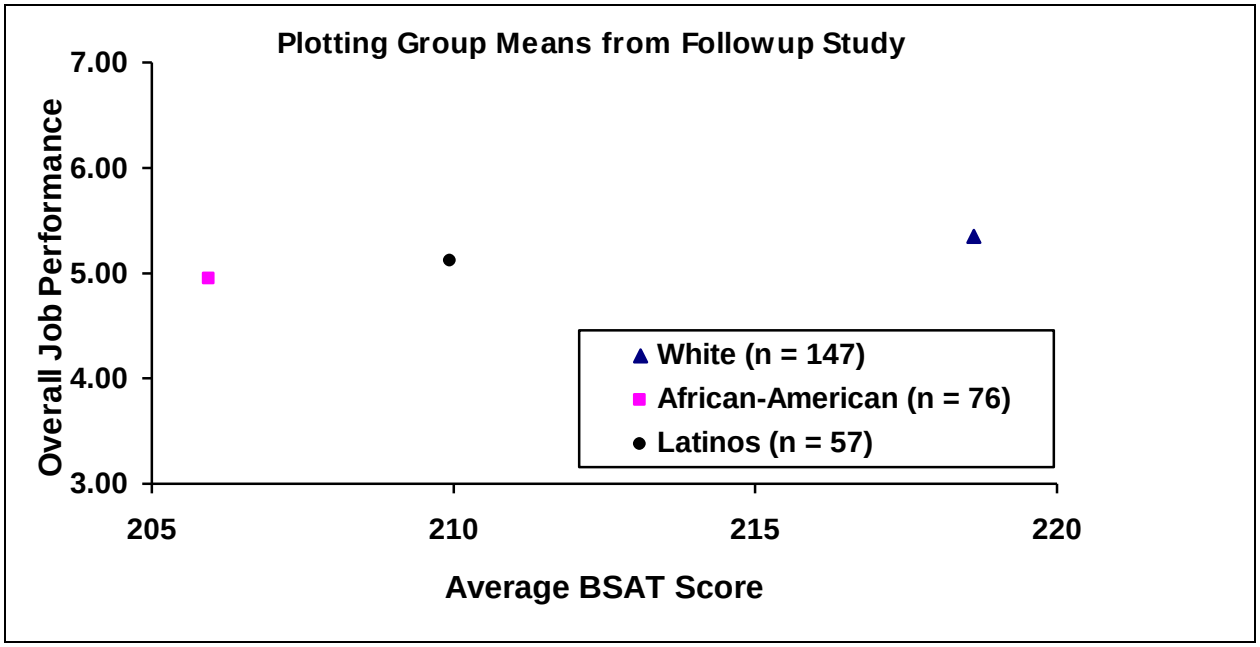


Figure 2: Scatterplotting BSAT score and overall job performance for the Preliminary and Followup Studies. Supervisors rated overall performance using a 9-point Likert-type scale in the Preliminary Study (1,2=exceptionally high; 3,4=very high; 5,6=moderately high; 7,8=moderately low; 9=unsatisfactory) and 7-point Likert-type scale in the Followup Study (1=exceptionable; 2=very high; 3=high; 4=average; 5=below average; 6=passable; 7=unacceptable). Scores on these rating scales have been reversed for ease of presentation.

2155-0182/10/$3.00

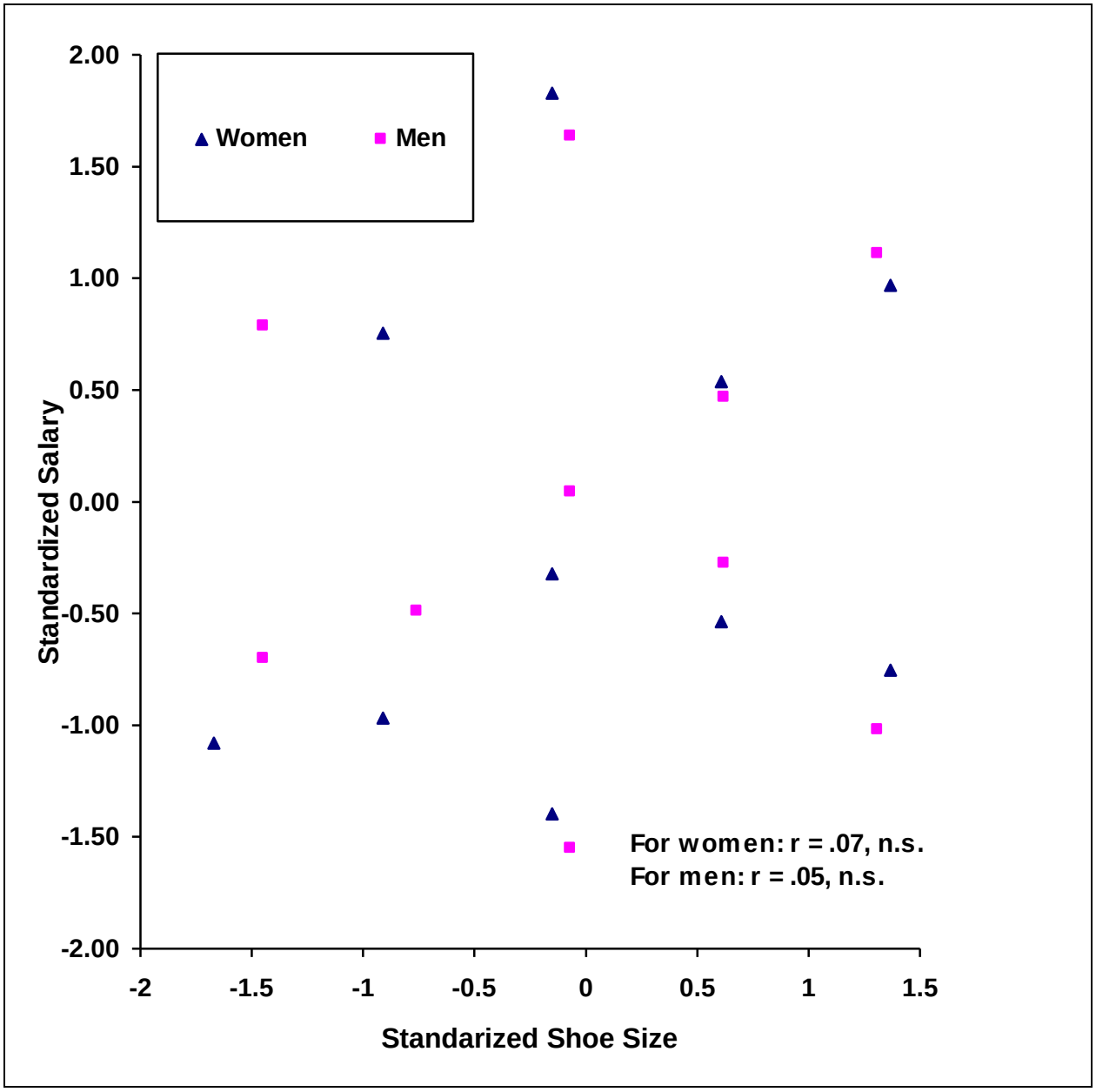


Figure 3: Correlating shoe size and salary using hypothetical data normatively standardized separately for women and men

Yet, typically researchers simply pool data across subgroups in total-sample analyses. This practice inflates total sample validities throughout the testing industry. Among the most robust findings in the literature on cognitive ability testing is that minorities score significantly lower on cognitive ability tests than do whites.[20] And in validation studies, minorities often receive significantly lower performance ratings.[21] Ironically, if test scores are lower for minorities than for whites, to meet the Uniform Guidelines' unfairness standard, minority performance ratings must also be lower. Although it is not unfair within the meaning of the Uniform Guidelines, this very situation will typically make tests appear more valid than they

2155-0182/10/$3.00

really are, if data are simplistically pooled and correlated. *Test developers should avoid indiscriminately pooling subgroup data, particularly when these subgroups have different means on the test and on the criterion*.

The Uniform Guidelines provide a basis for addressing the distortions arising from the improper pooling of data. Section 1607.14.B (4), entitled "Representativeness of the sample," relevantly provides: *Where samples are combined or compared*, attention should be given to see that such samples are comparable in terms of the actual job they perform, the length of time on the job where time on the job is likely to affect performance, and *other relevant factors likely to affect validity differences*; or that these factors are included in the design of the study and their effects identified (emphasis added).

Hardly restricted to industrial/organizational psychology, this false-correlation problem pervades the life sciences: indeed it has been stated that the problem of paradoxical confounding is the most significant and pervasive challenge to the validity of empirical quantitative analysis in all areas of inquiry.[22] The practice of simply pooling data across subgroups inflates correlation coefficients whenever one group has higher mean scores than the other on both x and y. For example, studies of naturalistic animal behavior often pool data across intact groups to examine relationships among social and behavioral variables, without regard to possible mean differences.[23] Similarly, personality psychologists often pool the data of males and females, examine the correlations among numerous measures of, for example, anxiety, neuroticism, and general maladjustment, and find a single, stable pervasive trait that they label negative affectivity.[24] Given that women tend to report higher levels of negative experience in general than do men[25], pooling male and female data without standardization will inflate the observed intercorrelations for the total group, exaggerating structural unidimensionality.

The problem of when and how to combine the data of multiple groups remains largely ignored in the social sciences.[16] Haphazardly pooling data across different groups (or time periods[16]) can produce unexpected, counterintuitive relationships, which researchers inevitably scramble to explain *a posteriori*. If one group scores lower than the other on x but higher on y, for example, then simply pooling the data across groups can produce a negative correlation for the total sample, even if the x-y relationship is actually positive in each group (the group with lower x scores and higher y scores will fall in the upper-left quadrant of the scatterplot, whereas the group with higher x scores and lower y scores will fall in the lower-right quadrant, yielding a false negative correlation). As a case in point, when studying psychosocial adjustment to head injury, researchers often combine the data of patients who are aware of functional deficits with the data of patients who are unaware of functional deficits. The correlation between severity of injury and emotional distress is then computed. An unexpected negative correlation often emerges, with greater severity of injury predictive of less distress.[26-28] It seems likely that the correlation between severity of injury and distress is actually positive within both the deficit-aware and deficit-unaware groups (i.e., greater severity linked to greater emotional distress), but that patients aware of their impairment have less severe head trauma (lower x-scores) and report higher levels of emotional distress (higher y-scores) than do patients who are unaware of their impairment, creating a false negative correlation for the pooled sample.

At first blush, the procedure of standardizing data separately within groups before computing pooled validity coefficients may seem similar to so-called *race norming*.[29] This latter practice seeks to ameliorate a test's adverse impact in personnel selection, by expressing individual test scores in terms of their standing relative to the mean of their particular racial group. However, the two approaches have entirely

2155-0182/10/$3.00

different objectives. Race-norming uses standardization in deciding which job applicants to hire. Standardizing raw data separately within groups before computing pooled validity coefficients, on the other hand, is done simply to avoid bias in estimating test validity, and is not used to select job applicants. Whereas race norming disaggregates data to avoid comparison between groups when selecting applicants, standardizing before computing pooled validity coefficients allows data from different groups to be meaningfully aggregated when evaluating test validity if their covariance is homogeneous.

Implications for Validity Generalization

Besides highlighting ambiguities in the Uniform Guidelines, the *Melendez* case also has implications for meta-analytic research on validity generalization.[10] This area of research entails synthesizing validity coefficients from studies attempting to validate personnel selection tests, in order to draw conclusions about the relationship between cognitive ability and job performance. Typically, these meta-analyses have concluded that cognitive ability tests are generally valid in the workplace across a full range of different racial subgroups, different jobs, different tests, and different settings.[10] Although conclusions about validity generalization have been criticized on a variety of statistical and conceptual grounds[30], the problem of paradoxical counfounding has been overlooked.

Validity coefficients based on pooled unstandardized data will be biased whenever the data contain subsamples that reliably differ on both the predictor and the criterion (e.g., racial subgroups, gender, types of jobs, different sites of data collection). Synthesizing validity coefficients will yield biased conclusions when the coefficients share a common bias (e.g., whites had higher test scores and higher performance ratings than other racial subgroups, and the data of racial subgroups were simply combined). This suggests that previous meta-analyses of test validation studies using total sample correlations have *overestimated* overall effect strength.

Although most statistical adjustments in meta-analysis serve to increase the strength of observed relationships by correcting for sources of unreliability[10], a comparable adjustment is needed to remove the inflation in correlations due to paradoxical confounding. If means and standard deviations are available for racial subgroups from the primary studies, for example, then group differences can be examined on the predictor (x) and the criterion (y). When one group scores higher than others on x or y, a better estimate of the pooled correlation coefficient is a weighted composite of the correlations for the separate subgroups, using *r*-to-*z* methodology.[18] Paradoxical confounding exists whenever the coefficient based on pooled data differs from the weighted mean coefficient across subgroups.

In the name of validity generalization, extravagant claims have been made for the efficacy of cognitive ability tests as personnel selection devices. For example, it has been argued: "[R]eliable measures of the standard aptitudes (e.g., verbal, quantitative, and spatial abilities) are valid predictors of... performance on the job for all jobs in the occupational spectrum... [T]hese findings can be generalized to all jobs in the economy for which tests are used in selection... [T]here are no jobs or job families for which reliable measures of cognitive ability do not have validity".[31] Couching claims in cosmic hyperbole, validity generalization is likened to "the powerful telescopes used in astronomy," and it is suggested that the theory is as well-established as the measurement of the speed of light.[31]

Ironically, persistent disparities between test scores and performance evaluations of majority and minority employees is also what one would expect from a pervasive pattern of discrimination. Consistent use of discriminatory employment tests, coupled with racially-biased supervisory evaluations, would produce com-

2155-0182/10/$3.00

parable statistical outcomes. For this result to obtain, overt and conscious racial discrimination need not exist. For example, unconscious, subjective perceptions favoring majority employees would tend to inflate the mean criterion measure for this subgroup; similarly, the impact of broad societal discrimination would tend to depress the mean test performance of a minority group. Where the data for such racial and ethnic groups are pooled without correcting for differences in means on predictor and criterion, the likely result is a distribution yielding false positive correlations. The resulting evidence of "validity" would be illusory.

The implications for the theory of validity generalization are clear. Meta-analysis is based in a vast pool of results from combined samples, drawn primarily from reported validity studies of employment tests. A systematic bias throughout this data base would correspondingly bias the meta-analysis. Further empirical research is needed to isolate and assess the statistical impact of artifactual validity arising from paradoxical confounding.

Conclusion

The case of *Melendez v. Illinois Bell Telephone Company* highlights ambiguities in the Uniform Guidelines for validating personnel selection tests. Although the Guidelines could be revised to clarify these ambiguities, there is a potential drawback to this approach: namely, the possibility that hard-won legal precedents, gained over the years in the courts, might be lost if the Guidelines were substantially modified.[30] There is an inevitable trade-off here between more specificity in the Uniform Guidelines, and less applicability of previous court rulings.

Although the judgment in the *Melendez* case strengthens the legal means for removing invalid, discriminatory tests from the workplace, it does not immediately reduce the likelihood of such tests being developed in the first place, as might revisions in the Uniform Guidelines. Ultimately, however, the demise of invalid discriminatory tests in the workplace may depend more on their perceived liability costs for the user than on the specificity of the guidelines for test development.

References

[1]*Melendez v. Illinois Bell Telephone Co*., No. 90 C 5020 (N.D. Ill. 1994); *aff'd*, 79 F.3d 661 (7th Cir. 1996).

[2]*Standards for educational and psychological testing*. Washington, DC, APA Books, 1985.

[3]*Principles for the validation and use of personnel selection procedures*. College Park, MD, Society for Industrial and Organizational Psychology, 1987.

[4]Adams E. Development of the BSAT qualification score for general management hires. Trenton, NJ, AT&T, 1982.

[5]Yarnold PR, Soltysik RC. *Optimal data analysis: a guidebook with software for Windows*. Washington, DC, APA Books, 2005.

[6]Rosenthal R, Rosnow RL. *Contrast analysis: focused comparisons in the analysis of variance*. London: Cambridge University Press, 1985.

[7]Cascio WF, Outtz J, Zedeck S, Goldstein IL. Statistical implications of six methods of test score use in personnel selection. *Human Performance* 1991, 4:233-264.

[8]Gottfredson LS. Reconsidering fairness: a matter of social and ethical priorities. *Journal of Vocational Behavior* 1988, 33:293-319.

[9]Hartigan JA, Wigdor AK. *Fairness in employment testing: validity generalization, minority issues, and the General Aptitude Test Battery*. Washington, DC, National Academy Press, 1989.

2155-0182/10/$3.00

[10]Hunter JL, Schmidt FL. Methods of meta-analysis. London: Sage, 1990.

[11]Peterson NS, Novick MR. An evaluation of some models for culture-fair selection. *Journal of Educational Measurement* 1976, 13:3-29.

[12]Tenopyr M. Fairness in employment testing. *Society* 1990, 27:17-20.

[13]Cleary TA. Test bias: Prediction of grades of negro and white students in integrated colleges. *Journal of Educational Measurement* 1968, 5: 115-124.

[14]Bryant FB, Yarnold PR. Principal-components analysis and exploratory and confirmatory factor analysis. In: LG Grimm, PR Yarnold (Eds.), *Reading and understanding multivariate statistics*. Washington, DC: American Psychological Association, 1995, pp. 99-136.

[15]Simpson RH. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society* 1951B, 13:238-241.

[16]Yarnold PR. Characterizing and circumventing Simpson's paradox for ordered bivariate data. *Educational and Psychological Measurment* 1996, 56:430-442.

[17]McClish DK. Combining and comparing area estimates across studies or strata. *Medical Decision Making* 1992, 12:274-279.

[18]Rosenthal, R. *Meta-analytic procedures for social research*. Beverly Hills, CA, Sage, 1984.

[19]Ghiselli, E.E. *The validity of occupational aptitude tests*. New York, Wiley, 1966.

[20]Wigdor, A.K., & Garner, W.R. *Ability testing: uses, consequences, and controversies: part 1. Report to the committee.* Washington, DC, National Academy Press, 1982.

[21]Miner MG, Miner JB. *Employee selection within the law*. Washington, DC, The Bureau of National Affairs, 1978.

[22]Soltysik RC, Yarnold PR. The use of unconfounded climatic data improves atmospheric prediction. *Optimal Data Analysis*, 1:67-100.

[23]Martin P, Bateson P. *Measuring behavior: an introductory guide*. Cambridge, England, Cambridge University Press, 1993.

[24]Watson D, Clark LA. Negative affectivity: the disposition to experience aversive emotional states. *Psychological Bulletin* 1984, 96:465-490.

[25]Gove WR, Tudor JF. Adult sex roles and mental illness. *American Journal of Sociology* 1973, 78:812-835.

[26]Landy PR. The post-traumatic syndrome in closed head injuries and accident neuroses. *Proceedings of the Australian Association of Neurology* 1968, 5:463-466.

[27]McClean A, Dikmen S, Temkin N, Wyler A, Gale JL. Psychosocial functioning at 1 month after head injury. *Neurosurgery* 1984, 14:393-399.

[28]Miller H. Accident neuroses: lectures I and II. *British Medical Journal* 1961, 1:919-952, 992-998.

[29]Gottfredson LS. The science and politics of race-norming. *American Psychologist* 1994, 49: 955-963.

[30]Seymour RT. Why plaintiffs' counsel challenge tests, and how they can successfully challenge the theory of "validity generalization." *Journal of Vocational Behavior* 1988, 33:331-364.

2155-0182/10/$3.00

Author Notes

The first author served as the plaintiff's expert witness and the second author as attorney for the plaintiff (Melendez) in the case of *Melendez v. Illinois Bell Telephone Company*. We wish to thank Robert Perloff, Richard Seymour, Charles Spielberger, and Scott Tindale for helpful comments on an earlier draft of this manuscript. Correspondence should be sent to Fred B. Bryant at: Department of Psychology, Loyola University Chicago, 6525 North Sheridan Road, Chicago, IL, 60626. Send e-mail to: fbryant@luc.edu.